Common-Message Broadcast Channels with Feedback in the Nonasymptotic Regime

Stop Feedback

Trillingsgaard, Kasper Fløe; Yang, Wei; Durisi, Giuseppe; Popovski, Petar

Published in:
I E E E Transactions on Information Theory

DOI (link to publication from Publisher):
10.1109/TIT.2018.2868953

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Common-Message Broadcast Channels with Feedback in the Nonasymptotic Regime: Stop Feedback

Kasper Fløe Trillingsgaard, *Member, IEEE*, Wei Yang, *Member, IEEE* , Giuseppe Durisi, *Senior Member, IEEE*, and Petar Popovski, *Fellow, IEEE*

***Abstract*—We investigate the maximum coding rate for a given average blocklength and error probability over a K-user discrete memoryless broadcast channel for the scenario where a common message is transmitted using variable-length stop-feedback codes. For the point-to-point case, Polyanskiy *et al.* (2011) demonstrated that variable-length coding combined with stop-feedback significantly increases the speed of convergence of the maximum coding rate to capacity. This speed-up manifests itself in the absence of a square-root penalty in the asymptotic expansion of the maximum coding rate for large blocklengths, i.e., *zero dispersion*. In this paper, we present nonasymptotic achievability and converse bounds on the maximum coding rate of the common-message K-user discrete memoryless broadcast channel, which strengthen and generalize the ones reported in Trillingsgaard *et al.* (2015) for the two-user case. An asymptotic analysis of these bounds reveals that zero dispersion cannot be achieved for certain common-message broadcast channels (e.g., the binary symmetric broadcast channel). Furthermore, we identify conditions under which our converse and achievability bounds are tight up to the second order. Through numerical evaluations, we illustrate that our second-order expansions approximate accurately the maximum coding rate and that the speed of convergence to capacity is indeed slower than for the point-to-point case.**

***Index Terms*—Broadcast channel with common-message, finite blocklength regime, stop feedback, decision feedback, channel dispersion, variable-length coding.**

I. Introduction

We consider the setup in which an encoder wishes to convey a common message over a discrete memoryless broadcast channel with feedback from K decoders. Similarly to the single-decoder case, *full feedback* (i.e., instantaneous feedback of the received symbols) combined with *fixed-blocklength codes* does not improve capacity, which is given by [2, p. 126]

$$C = \sup_{P} \min_{k\in\{1,\cdots,K\}} I(P, W_k). \tag{1}$$

Here, $W_1, \cdots, W_K$ denote the channels to the decoders $1, \ldots, K$, respectively, and the supremum is over all input distributions P. For the case of no feedback, the common-message broadcast channel is equivalent to a compound channel, and the speed at which C is approached as the blocklength n increases is of the order $1/\sqrt{n}$ (see [3]), which is the same as in the single-decoder (point-to-point) case [4]. Specifically, the logarithm of the maximum number of codewords $M^*(n,\epsilon)$ that can be transmitted with blocklength n and maximum error probability ϵ can be expanded as [3]

$$\frac{1}{n}\log M^*(n,\epsilon) = C - \sqrt{\frac{V_{\text{no-fb}}}{n}}Q^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}}\right) \tag{2}$$

where

$$\sqrt{V_{\text{no-fb}}} = \min_{\mathbf{v}:\sum_x v_x=0} \max_{k\in\{1,\ldots,K\}} \left\{\nabla I_k(\mathbf{v}) + \sqrt{V_k}\right\}. \tag{3}$$

Here, V_k denotes the conditional information variance of component channel k evaluated at the unique capacity-achieving distribution P^* (see (9)) and $\nabla I_k(\mathbf{v})$ denotes the directional derivative of the mutual information of decoder k at P^* (see (12)).

For point-to-point channels, although feedback does not increase capacity, it improves dramatically the error exponent, provided that *variable-length codes* are used. This was first demonstrated by Burnashev who found that the error exponent for this setting is given by [5]

$$E(R) = \frac{\widetilde{C}_1}{\widetilde{C}}(\widetilde{C} - R) \tag{4}$$

for all rates $0 < R < \widetilde{C}$. Here, $\widetilde{C}$ denotes the channel capacity for the point-to-point case and $\widetilde{C}_1$ denotes the maximum relative entropy between two conditional output distributions. Yamamoto and Itoh [6] proposed a two-phase scheme that attains the error exponent in (4). Furthermore, Berlin *et al.* [7] provided an alternative and simpler converse proof to (4), which parallels the two-phase scheme proposed in [6].

In the fixed-error regime, Polyanskiy *et al*. [8] found that the speed at which the maximum coding rate converges to capacity

The work of K. F. Trillingsgaard and P. Popovski was supported by the European Research Council (ERC Consolidator Grant Nr. 648382 WILLOW) within the Horizon 2020 Program. The work of G. Durisi was supported by the Swedish Research Council under the grant 2016-032931. The material of this paper was presented in part at the 2016 IEEE International Symposium on Information Theory [1].

K. F. Trillingsgaard and P. Popovski are with the Department of Eletronic Systems, Aalborg University, 9220, Aalborg Øst, Denmark (e-mail: {kft,petarp}@es.aau.dk).

W. Yang is with Qualcomm Technologies, Inc., San Diego, 92121, USA (e-mail: weiyang@qti.qualcomm.com).

G. Durisi is with the Department of Electrical Engineering, Chalmers University of Technology, 41296, Gothenburg, Sweden (e-mail: durisi@chalmers.se).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

is significantly improved in the presence of full feedback and variable-length codes. Specifically, they showed that

$$\frac{1}{\ell}\log \widetilde{M}_{\rm f}^*(\ell,\epsilon) = \frac{\widetilde{C}}{1-\epsilon} - \mathcal{O}\left(\frac{\log \ell}{\ell}\right) \tag{5}$$

where ℓ denotes the average blocklength (average transmission time) and $\widetilde{M}_{\rm f}^*(\ell,\epsilon)$ is the maximum number of codewords that can be transmitted with average transmission time ℓ and average error probability ϵ in the point-to-point case. One sees from (5) that no square-root penalty occurs (*zero dispersion*), which implies a fast convergence to the asymptotic limit. This fast convergence is demonstrated numerically in [8] by means of nonasymptotic bounds.

When fixed-length codes are used, it has been shown in [8] and [9] that feedback does not improve the second-order term in the large-blocklength expansion of the maximum number of codewords for a large class of channels with certain symmetry properties. However, for some channels with nonunique capacity-achieving input distributions [9] feedback results in a larger second-order term.

In this paper, we shall be concerned with the scenario in which the feedback channel is only used to stop transmissions (stop/decision feedback). Following [8], we shall refer to variable-length coding schemes relying on stop feedback as variable-length stop-feedback (VLSF) codes. It was shown in [8], [10], [11] that the error exponent

$$E(R) = \widetilde{C} - R \tag{6}$$

is achievable using VLSF codes. However, the tightest converse bound known is the full-feedback error exponent (4). Stop feedback is sufficient to achieve the zero-dispersion result in (5). However, also in this case, the tightest nonasymptotic converse bound available for VLSF codes is the full-feedback converse reported in [8].[1]

When only stop feedback is available, the zero-dispersion result (5) does not necessarily hold for the common-message discrete memoryless broadcast channels (CM-DMBC) considered in this paper. Specifically, we showed in [13] that there exist CM-DMBCs for which the second term in the asymptotic expansion of the maximum coding rate achievable with VLSF codes is of order $1/\sqrt{\ell}$ (cf. (5)). Our analysis in [13] is limited to the two-user case and relies on the restrictive assumption that there exists a unique input distribution P^* that simultaneously maximizes $I(P,W_1)$ and $I(P,W_2)$. Furthermore, the upper and lower bounds on the maximum coding rate provided in [13] do not match up to the second order. In this paper, we refine the results obtained in [13] and extend them to a broader class of common-message broadcast channels.

Contribution: Focusing on VLSF codes, we obtain nonasymptotic achievability and converse bounds on the maximum number of codewords $M_{\rm sf}^*(\ell,\epsilon)$ with average blocklength ℓ that can be transmitted on a CM-DMBC with reliability $1-\epsilon$. Here, the subscript "sf" stands for stop feedback. By analyzing these bounds in the large-ℓ regime, we prove that when the K component channels are independent (in the sense made precise in (7)) and when the mutual information evaluated at the capacity-achieving input distribution equals C for two or more component channels, then the asymptotic expansion of $\log M_{\rm sf}^*(\ell,\epsilon)$ contains a square-root penalty, provided that some mild technical conditions are satisfied. Thus, we cannot expect the same fast convergence to capacity as in the point-to-point case. The intuition behind this result is as follows: in the point-to-point case, the stochastic overshoots of the information density that result in the square-root penalty can be virtually eliminated by using variable-length coding with stop-feedback. Indeed, decoding is stopped after the information density exceeds a certain threshold, which yields only negligible stochastic variations. In the multiuser setup, however, the stochastic variations in the difference between the stopping times at the decoders make the square-root penalty reappear. Note that our result does not necessarily imply that feedback is useless. It only shows that VLSF codes cannot be used to speed-up convergence to the same level as in the point-to-point case. We also obtain upper and lower bounds on the second-order term in the asymptotic expansion of $\log M_{\rm sf}^*(\ell,\epsilon)$ that generalize and tighten the ones reported in [13]. The bounds turn out to match in certain special cases, e.g., when, in a two-user case, P^* simultaneously maximizes $I(P,W_1)$ and $I(P,W_2)$ (the case treated in [13]). Numerical evaluations of our nonasymptotic achievability and converse bounds reveal that the asymptotic expansion of $\log M_{\rm sf}^*(\ell,\epsilon)$ obtained in this paper yields an accurate approximation for the maximum coding rate.

We remark that many of the results in this paper first appeared in the conference paper [1]. The present paper includes proofs that were omitted in [1] as well as several intuitive remarks, discussion, and additional numerical results.

Notation: We denote the n-dimensional all-zero vector and the n-dimensional all-one vector by $\mathbf{0}_n$ and $\mathbf{1}_n$, respectively. Vectors are denoted by boldface letters (e.g., $\mathbf{x}$), while their entries are denoted by roman letters (e.g., x_i). The length of a vector is denoted by $\mathrm{len}(\cdot)$ and the Euclidean norm by $\|\cdot\|$. For a differentiable function $f(\cdot)$, we let $f'(\cdot)$ denote its derivative. Upper case, lower case, and calligraphic letters indicate random variables (RV), deterministic quantities, and sets, respectively. The cardinality of a set is denoted by $|\cdot|$ (e.g., $|\mathcal{A}|$). We let x_m^n denote the tuple $(x_m,\cdots,x_n)$. For the channel outputs at decoder k, we let $y_{k,m}^n$ denote the tuple $(y_{k,m},\ldots,y_{k,n})$. When $m=1$, the subscript is sometimes omitted. We denote the set of probability distributions on $\mathcal{A}$ by $\mathcal{P}(\mathcal{A})$ and the support of a probability mass function P by $\mathrm{supp}(P)$. For a RV X with probability distribution P, we let P^n denote the joint probability distribution of the vector $[X_1,\cdots,X_n]$, where $\{X_i\}$ are independently and identically distributed (i.i.d.) according to P. The probability density function of a standard Gaussian RV is denoted by $\phi(\cdot)$. Furthermore, $\Phi(x) \triangleq 1-Q(x)$ is its cumulative distribution function, with $Q(\cdot)$ being the Q function. We let x^+ denote $\max(0,x)$. Throughout the paper, $\log(\cdot)$ is the base e logarithm and the index k always belongs to the set $\mathcal{K} \triangleq \{1,\cdots,K\}$, although this is sometimes not explicitly mentioned. We use "$\mathbb{c}$" to denote a finite nonnegative constant. Its value may change at each occurrence. For two functions $f(\cdot)$ and $g(\cdot)$, the notation $f(x)=\mathcal{O}(g(x))$, as $x\to\infty$, means that $\limsup_{x\to\infty}|f(x)/g(x)|<\infty$, $f(x)=o(g(x))$, as $x\to\infty$,

[1] An exception is the binary erasure channel, for which a nonasymptotic converse bound for the case of stop feedback that is tighter than the ones for full feedback is reported in [12].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

means that $\lim_{x\to\infty}|f(x)/g(x)| = 0$, and $f(x) = \Theta(g(x))$, as $x \to \infty$, means that $c \le f(x)/g(x) \le C$ for two positive constants c and C, $c < C$, and for all sufficiently large x. Finally, $\mathbb{N}$ denotes the set of positive integers, $\mathbb{Z}_+ \triangleq \mathbb{N} \cup \{0\}$, the symbol $\mathbb{R}$ indicate the set of real numbers, and $\mathbb{R}_0^n$ denotes the set $\left\{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\right\}$.

II. System Model

A CM-DMBC with K decoders consists of a finite-cardinality input alphabet $\mathcal{X}$, and finite-cardinality output alphabets $\{\mathcal{Y}_k\}$, along with K stochastic matrices $\{W_k\}$, where $W_k(y_k|x)$ denotes the probability that $y_k \in \mathcal{Y}_k$ is observed at decoder k given the channel input $x \in \mathcal{X}$. We assume, without loss of generality, that $\mathcal{X} = \{1, \cdots, |\mathcal{X}|\}$. The outputs at time t are assumed to be conditionally independent given the input, i.e.,

$$P_{Y_{1,t},\cdots,Y_{K,t}|X_t}(y_{1,t},\cdots,y_{K,t}|x_t) \triangleq \prod_k W_k(y_{k,t}|x_t). \quad (7)$$

Let $P \times W_k : (x, y_k) \mapsto P(x)W_k(y_k|x)$ denote the joint probability distribution of input and output at decoder k. Finally, let $PW_k : y_k \mapsto \sum_{x\in\mathcal{X}} P(x)W_k(y_k|x)$ denote the induced marginal distribution on $\mathcal{Y}_k$. For every $P \in \mathcal{P}(\mathcal{X})$ and $n \in \mathbb{N}$, the information density is defined as

$$i_{P,W_k}(x^n; y_k^n) \triangleq \sum_{i=1}^n \log \frac{W_k(y_{k,i}|x_i)}{PW_k(y_{k,i})}. \quad (8)$$

We let $I_k(P) \triangleq \mathbb{E}_{P\times W_k}[i_{P,W_k}(X; Y_k)]$ be the mutual information,

$$V_k(P) \triangleq \mathbb{E}_P[\mathrm{Var}_{P\times W_k}[i_{P,W_k}(X; Y_k)|X]] \quad (9)$$

be the conditional information variance, and

$$T_k(P) \triangleq \mathbb{E}_{P\times W_k}\left[|i_{P,W_k}(X; Y_k) - I_k(P)|^3\right] \quad (10)$$

be the third absolute moment of the information density. Here, $\mathbb{E}_{P\times W_k}[\cdot]$ and $\mathrm{Var}_{P\times W_k}[\cdot]$ denote the expectation and the variance, respectively, when the joint probability distribution of (X, Y_k) is $P \times W_k$, and $\mathbb{E}_P[\cdot]$ denotes the expectation when the probability distribution on X is P. The capacity of the CM-DMBC is given by (1), where the supremum is over all probability distributions $P \in \mathcal{P}(\mathcal{X})$. We restrict ourselves to the case where the supremum in (1) is attained by a unique probability distribution P^*. The corresponding (unique) capacity-achieving output distribution for decoder k is denoted by $P^*_{Y_k}$. Furthermore, the individual capacities of each of the discrete memoryless component channels $\{W_k\}$ are denoted by

$$C_k \triangleq \sup_{P\in\mathcal{P}(\mathcal{X})} I_k(P). \quad (11)$$

Finally, we let $V_k \triangleq V_k(P^*)$ and let $\nabla I_k(\mathbf{v})$ denote the directional derivative of the mutual information $I_k(P)$ along the direction $\mathbf{v} \in \mathbb{R}_0^{|\mathcal{X}|}$ at the point P^*

$$\nabla I_k(\mathbf{v}) \triangleq \sum_{x\in\mathcal{X}} v_x D(W_k(\cdot|x)||P^*_{Y_k}). \quad (12)$$

Here, $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

In addition to (7) and to the uniqueness of P^*, we shall also assume that the channel laws $\{W_k\}$ satisfy the following conditions:

1) $I_k(P^*) = C$ for every $k \in \mathcal{K}$.
2) $V_k(P^*) > 0$ for every $k \in \mathcal{K}$.
3) $P^*(x) > 0$ for all $x \in \mathcal{X}$.

The first condition is not critical, and it is added only to simplify the statement of our results. Indeed, the nonasymptotic bounds we shall present in Theorem 1 and 3 also hold for the case when $I_k(P^*) > C$ for some k. Furthermore, our positive dispersion result (Theorem 5) also holds when the first condition is violated, provided that there exist at least two component channels k_1 and k_2 such that $I_{k_1}(P^*) = I_{k_2}(P^*) = C$. This is because the decoders whose component channels satisfy $I_k(P^*) > C$ feed back their stop signals much earlier than the remaining decoders and therefore do not contribute to the asymptotic expansion of $\log M^*_{\mathrm{sf}}(\ell, \epsilon)$. If $I_k(P^*) = C$ for a single component channel, then zero dispersion can be attained. This may happen in certain practical scenarios, e.g., when all receivers are at different distances from the transmitter.

We are now ready to formally define a VLSF code for the CM-DMBC.

Definition 1: An (ℓ, M, ϵ)-VLSF code for the CM-DMBC consists of:

1) A RV $U \in \mathcal{U}$, with $|\mathcal{U}| \le K + 1$, which is known at both the encoder and the decoders.[2]
2) A sequence of encoders $f_n : \mathcal{U}\times\mathcal{M} \mapsto \mathcal{X}$, each one mapping the message J, drawn uniformly at random from the set $\mathcal{M} \triangleq \{1, \ldots, M\}$, to the channel input $X_n = f_n(U, J)$.
3) Nonnegative integer-valued RVs $\tau_1, \cdots, \tau_K$ that are stopping times with respect to the filtrations (see [14, p. 488]) $\mathcal{F}_{k,n} \triangleq \sigma\{U, Y_k^n\}$ and satisfy
$$\mathbb{E}\left[\max_k \tau_k\right] \le \ell. \quad (13)$$
4) A sequence of decoders $g_{k,n} : \mathcal{U} \times \mathcal{Y}_k^n \mapsto \mathcal{M}$ satisfying
$$\mathbb{P}\left[J \ne g_{k,\tau_k}(U, Y_k^{\tau_k})\right] \le \epsilon, \qquad k \in \mathcal{K}. \quad (14)$$

The maximum number of codewords with average length ℓ and error probability not exceeding ϵ is denoted by

$$M^*_{\mathrm{sf}}(\ell, \epsilon) \triangleq \max\{M : \exists(\ell, M, \epsilon)\text{-VLSF code}\}. \quad (15)$$

Some remarks on Definition 1 are in order. VLSF codes require a feedback link from each decoder to the encoder. This feedback consists of a 1-bit "stop signal" per decoder which is sent by decoder k at time τ_k. The encoder continuously transmits until all decoders have fed back a stop signal. Hence, the blocklength is $\max_k \tau_k$. Note also that, differently from the full-feedback case, the encoder output at time n depends on the message and on the common randomness U, but does not depend on the past output signals $\{Y_k^{n-1}\}$ and it is also independent of the stop signals received before time n. The stop signals are also only available at the encoder but not at the other decoders. This implies that τ_k and g_{k,τ_k} depend only on the common randomness U and on the output sequence $Y_k^{\tau_k}$, but does not depend on $\{\tau_{k'}\}_{k'\ne k}$. Allowing the encoder output at time n to depend on the previously

[2]Some remarks on the role of U can be found after this definition.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

received stop signal may yield to a faster convergence to capacity than what reported in this paper. From a practical perspective, however, this dependency complicates the design of the encoder and the decoders. Specifically, the encoder may need to use multiple codebooks depending on which stop signals are received and the decoders need to detect when the encoder switches between codebooks.

Note also that our definition of average blocklength (13) is inherently "encoder-centric". An alternative, decoder-centric approach would be to require that $\max_k \mathbb{E}[\tau_k] \leq \ell$. Under such an alternative definition, the zero-dispersion result from [8] continues to hold.

The RV U serves as common randomness between the transmitter and all receivers, and enables the use of randomized codes [15]. As for the proof of [8, Th. 3], randomized codes are necessary to prove our achievability bound. This is because we need to prove the existence of a code simultaneously satisfying (13) and (14). Note that the classic random-coding argument would allow us to establish the existence of a deterministic VLSF-code (a VLSF-code with $|\mathcal{U}| = 1$) satisfying only one of the constraints. To establish the bound on the cardinality of U provided in Definition 1, one can proceed as in [8, Th. 19] and use Caratheodory theorem to show that $|\mathcal{U}| \leq K + 2$. This bound can be further improved to $|\mathcal{U}| \leq K + 1$ by using the Fenchel-Eggleston theorem [16, p. 35] in place of Caratheodory theorem.

III. Main Results

A. Nonasymptotic Achievability Bound

We provide below a K-user generalization of the nonasymptotic achievability bound reported in [13, Th. 1].[3]

Theorem 1: Fix a probability distribution P_{X^∞} on $\mathcal{X}^\infty$. Let $\gamma \geq 0$ and $0 \leq q \leq 1$ be arbitrary scalars. Let the joint probability distribution of $(X^n, \bar{X}^n, Y_1^n, \cdots, Y_K^n)$ be

$$\begin{aligned} &P_{X^n,\bar{X}^n,Y_1^n,\cdots,Y_K^n}(x^n,\bar{x}^n,y_1^n,\cdots,y_K^n) \\ &= P_{Y_1^n,\cdots,Y_K^n|X^n}(y_1^n,\cdots,y_K^n|x^n) P_{X^n}(x^n) P_{X^n}(\bar{x}^n) \end{aligned} \tag{16}$$

for all $n \in \mathbb{Z}_+$ and define the stopping times $\tau_k^{(0)}$ and $\bar{\tau}_k^{(0)}$, $k \in \mathcal{K}$, as follows:

$$\tau_k^{(0)} \triangleq \inf\left\{n \geq 0 : i_{P_{X^n},W_k^n}(X^n; Y_k^n) \geq \gamma\right\} \tag{17}$$

$$\bar{\tau}_k^{(0)} \triangleq \inf\left\{n \geq 0 : i_{P_{X^n},W_k^n}(\bar{X}^n; Y_k^n) \geq \gamma\right\}. \tag{18}$$

For every M, there exists an (ℓ, M, ϵ)-VLSF code such that

$$\ell \leq (1-q)\mathbb{E}\left[\max_k \tau_k^{0)}\right] \tag{19}$$

$$\epsilon \leq \max_k \left\{q + (1-q)(M-1)\mathbb{P}\left[\tau_k^{(0)} \geq \bar{\tau}_k^{(0)}\right]\right\} \tag{20}$$

$$\leq q + (1-q)(M-1)\exp\{-\gamma\}. \tag{21}$$

Proof: The proof of Theorem 1 follows closely the proof of [8, Th. 3]. See Appendix I for details. ∎

If the constant q is set to 0, then we obtain a straightforward generalization of [8, Th. 3]. The constant q in Theorem 1 is used to enable time-sharing. With probability q, the decoders simultaneously send stop signals to the encoder at time 0. The common randomness U can be used to enable this weak form of cooperation among the decoders.

[3] Note that there is a typo in [13, Eq. (12)]: a maximization over k is missing.

B. Nonasymptotic Converse Bound

Let $\mathscr{Y}_k$ denote all possible sequences (of arbitrary length) of symbols from $\mathcal{Y}_k$, i.e., $\mathscr{Y}_k \triangleq \{[\,]\} \cup \bigcup_{n=1}^\infty \mathcal{Y}_k^n$, where $[\,]$ denotes the vector of length 0. A subset $\overline{\mathscr{Y}}_k$ of $\mathscr{Y}_k$ is called complete prefix-free if and only if, for every $\mathbf{y} \in \mathcal{Y}_k^\infty$, there exists a unique $\bar{\mathbf{y}} \in \overline{\mathscr{Y}}_k$ such that $\bar{\mathbf{y}}$ is a prefix to $\mathbf{y}$, i.e., $\bar{\mathbf{y}} = [y_1, \cdots, y_{\text{len}(\bar{\mathbf{y}})}]$. The role of the complete prefix-free subsets of $\mathscr{Y}_k$ is to provide an equivalent representation of the stopping time τ_k. Indeed, given a stopping time τ_k, there exists a complete prefix-free subset $\overline{\mathscr{Y}}_k^{(u)}$ of $\mathscr{Y}_k$ for each $u \in \mathcal{U}$ such that $Y_k^{\tau_k} \in \overline{\mathscr{Y}}_k^{(U)}$. Conversely, every set of complete prefix-free subsets $\{\overline{\mathscr{Y}}_k^{(u)}\}_{u\in\mathcal{U}}$ also defines a stopping time $\tau_k = \min\{n \in \mathbb{Z}_+ : Y_k^n \in \overline{\mathscr{Y}}_k^{(U)}\}$, i.e., τ_k is a RV that depends only on the realizations of U and of Y_k^∞. Let $Q_k^{(\infty)}$ be an arbitrary probability measure on $\mathscr{Y}_k$ and define the mapping $Q_k : \mathscr{Y}_k \mapsto [0,1]$ as follows:

$$Q_k(\bar{\mathbf{y}}) \triangleq \sum_{\substack{\mathbf{y}\in\mathcal{Y}_k^\infty: \\ [y_1,\cdots,y_{\text{len}(\bar{y})}]=\bar{\mathbf{y}}}} Q_k^{(\infty)}(\mathbf{y}), \qquad \bar{\mathbf{y}} \in \mathscr{Y}_k. \tag{22}$$

We shall use the convention that $[y_1, \cdots, y_{\text{len}(\bar{\mathbf{y}})}] = [\,]$ when $\text{len}(\bar{\mathbf{y}}) = 0$. For every complete prefix-free subset $\overline{\mathscr{Y}}_k \subset \mathscr{Y}_k$, we observe that $Q_k(\cdot)$ defines a probability measure on $\overline{\mathscr{Y}}_k$. Indeed,

$$1 = \sum_{\mathbf{y}\in\mathcal{Y}_k^\infty} Q_k^{(\infty)}(\mathbf{y}) \tag{23}$$

$$= \sum_{\bar{\mathbf{y}}\in\overline{\mathscr{Y}}_k} \sum_{\substack{\mathbf{y}\in\mathcal{Y}_k^\infty: \\ [y_1,\cdots,y_{\text{len}(\bar{\mathbf{y}})}]=\bar{\mathbf{y}}}} Q_k^{(\infty)}(\mathbf{y}) \tag{24}$$

$$= \sum_{\bar{\mathbf{y}}\in\overline{\mathscr{Y}}_k} Q_k(\bar{\mathbf{y}}). \tag{25}$$

Based on $Q_k(\cdot)$, we define the *log-likelihood ratio*

$$i_k(x^n; y_k^n) \triangleq \log \frac{P_{Y_k^n|X^n}(y_k^n|x^n)}{Q_k(y_k^n)} \tag{26}$$

for $x^n \in \mathcal{X}^n$, $y_k^n \in \mathcal{Y}^n$, and $n \in \mathbb{N}$ with the convention that $i_k([\,];[\,]) = 0$.

To prove our nonasymptotic converse bound, we shall make use of the following lemma, which provides an information spectrum-type converse for VLSF codes. The proof of this lemma relies on a non-standard application of the meta-converse theorem [4, Th. 26]. Specifically, the meta-converse is applied to a general channel whose channel inputs are infinite-dimensional vectors and whose channel outputs are variable-length vectors belonging to a complete prefix-free subset of $\mathscr{Y}_k$.

Lemma 2: Fix arbitrary probability measures $Q_k^{(\infty)}$ on $\mathscr{Y}_k$, an arbitrary constant $\eta > 0$, and an (ℓ, M, ϵ)-VLSF code whose encoders induce a conditional probability distribution $P_{\mathbf{X}}^{(u)}$ on $\mathcal{X}^\infty$ given $U = u$ and whose stopping times are equivalently defined by the complete prefix-free subsets $\{\overline{\mathscr{Y}}_k^{(u)}\}_{u\in\mathcal{U}}$ of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

set $\mathscr{Y}_k$. There exist positive constants $\varepsilon_k^{(u)}$, defined for all $u \in \mathcal{U}$, and satisfying $\mathbb{E}_U\left[\varepsilon_k^{(U)}\right] \leq \epsilon + \eta$, such that

$$\mathbb{P}^{(u)}\Big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) < \log(\eta M)\Big] \leq \varepsilon_k^{(u)} \tag{27}$$

for every $\bar{\mathbf{x}} \in \mathrm{supp}(P_{\mathbf{X}}^{(u)})$ and $k \in \mathcal{K}$. The probability measure on $\mathcal{X}^\infty \times \overline{\mathscr{Y}}_1^{(u)} \times \cdots \times \overline{\mathscr{Y}}_K^{(u)}$ required to evaluate (27) is

$$\mathbb{P}^{(u)}_{\mathbf{X},\overline{\mathbf{Y}}_1,\cdots,\overline{\mathbf{Y}}_K}(\mathbf{x}, \bar{\mathbf{y}}_1, \cdots, \bar{\mathbf{y}}_K) \triangleq P_{\mathbf{X}}^{(u)}(\mathbf{x}) \prod_{k=1}^{K} \prod_{i=1}^{\mathrm{len}(\bar{\mathbf{y}}_k)} W_k(\bar{y}_{k,i}|x_i). \tag{28}$$

Here, we use the convention that $\prod_{i=1}^{0} W_k(\bar{y}_{k,i}|x_i) = 1$.

Proof: See Appendix II. ∎

We are now ready to state and prove our converse bound, which provides us with a lower bound on the average blocklength given M, an arbitrary probability measure $Q_k^{(\infty)}$, and an arbitrary positive constant η.

Theorem 3: For arbitrary probability measures $Q_k^{(\infty)}$ on $\mathscr{Y}_k$, and arbitrary $M \in \mathbb{N}$, $t \in \mathbb{Z}_+$, $\eta > 0$, and $\varepsilon_k \in (0,1)$, $k \in \mathcal{K}$, define the following function:[4]

$$L_t(\varepsilon_1, \cdots, \varepsilon_K) \triangleq \max_{x^t \in \mathcal{X}^t} \prod_k \min\Bigg\{1, \mathbb{P}\Bigg[\max_{0\leq n\leq t} i_k(x^n; Y_k^n) \geq \log M + \log\eta\Bigg] + \varepsilon_k\Bigg\}. \tag{29}$$

Here, the vector x^n contains the first n entries of x^t, and $Y_k^t \sim P_{Y_k^t|X^t=x^t}$. Then, every (ℓ, M, ϵ)-VLSF code must satisfy

$$\ell \geq \min_{\substack{P_U \in \mathcal{P}(\mathcal{U}), \varepsilon_k^{(u)} \in [0,1]: \\ \mathbb{E}_U\left[\varepsilon_k^{(U)}\right] \leq \epsilon+\eta}} \mathbb{E}_U\left[\sum_{t=0}^{\infty}\Big(1 - L_t\Big(\varepsilon_1^{(U)}, \cdots, \varepsilon_K^{(U)}\Big)\Big)\right]. \tag{30}$$

Proof: To establish Theorem 3, we derive a lower bound on the average blocklength ℓ that holds for all VLSF codes having M codewords and error probability no larger than ϵ. Fix an arbitrary (ℓ, M, ϵ)-VLSF code. It follows from (13) and from the conditional independence of the stopping times $\{\tau_k\}$ given U and $\mathbf{X}$ that

$$\ell \geq \mathbb{E}\Big[\mathbb{E}\Big[\max_k \tau_k | U, \mathbf{X}\Big]\Big] \tag{31}$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty}\Big(1 - \mathbb{P}\Big[\max_k \tau_k \leq t | U, \mathbf{X}\Big]\Big)\right] \tag{32}$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty}\left(1 - \prod_k \mathbb{P}^{(U)}\Big[\mathrm{len}\big(\overline{\mathbf{Y}}_k\big) \leq t | \mathbf{X}\Big]\right)\right]. \tag{33}$$

Here, we have used that $\tau_k = \mathrm{len}\big(\overline{\mathbf{Y}}_k\big)$. Hence, we can lower-bound ℓ by upper-bounding $\mathbb{P}^{(u)}\Big[\mathrm{len}\big(\overline{\mathbf{Y}}_k\big) \leq t | \mathbf{X} = \mathbf{x}\Big]$ for every $t \in \mathbb{Z}_+$.

[4]As clarified in the proof of the theorem, for a fixed $U = u$ and a set of error probabilities $\varepsilon_k^{(u)}$ for the decoders, the function $L_t\Big(\varepsilon_1^{(u)}, \cdots, \varepsilon_K^{(u)}\Big)$ is an upper bound on the conditional probability that $\max_k \tau_k \leq t$ given $U = u$.

Now, set

$$\varepsilon_k^{(u)}(\bar{\mathbf{x}}) \triangleq \mathbb{P}^{(u)}\Big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) < \lambda \Big| \mathbf{X} = \bar{\mathbf{x}}\Big]. \tag{34}$$

Then, it follows by Lemma 2 that we must have

$$\mathbb{E}_{\mathbf{X},U}\left[\varepsilon_k^{(U)}(\bar{\mathbf{X}})\right] \leq \epsilon + \eta. \tag{35}$$

From an intuitive perspective, (35) serves as a constraint on the stopping times. Namely, given M, k, and $\{\varepsilon_k^{(u)}(\bar{\mathbf{x}})\}$, the information density must exceed a threshold with probability larger than or equal $1 - \varepsilon_k^{(u)}(\bar{\mathbf{x}})$ when the stop signals are sent. Note also that (35) depends on the choice of $Q_k^{(\infty)}$ through the information density and on $\{\tau_k\}$ through $\{\overline{\mathbf{Y}}_k\}$.

Next, we upper-bound $\mathbb{P}^{(U)}\Big[\mathrm{len}\big(\overline{\mathbf{Y}}_k\big) \leq t | \mathbf{X} = \bar{\mathbf{x}}\Big]$ for every $u \in \mathcal{U}$ and $\bar{\mathbf{x}} \in \mathrm{supp}(P_{\mathbf{X}}^{(u)})$. Since the stopping times $\{\tau_k\}$ are conditionally independent given $U = u$ and $\mathbf{X} = \bar{\mathbf{x}}$, we have the steps (36)–(41), shown in the top of the next page. Here, (39) follows from (35); in (40), we let Y_k^n be distributed according to $P_{Y_k^n|X^n=x^n}$; finally, (41) follows from (29). Note that the probability term in (40) does not depend on the code.

Roughly speaking, the steps (36)–(40) dispose of the dependency on $\{\tau_k\}$. The intuition behind these steps is as follows: First, define the auxiliary stopping times

$$\hat{\tau}_k \triangleq \begin{cases} \min\{n : \imath_k(\mathbf{X}; Y_k^n) > \lambda\} & \text{if } \max_{0\leq n\leq t} \imath_k(\mathbf{X}; Y_k^n) > \lambda \\ t & \text{if } \max_{0\leq n\leq t} \imath_k(\mathbf{X}; Y_k^n) \leq \lambda \\ & \quad \text{or } A_k^{(U)}(\bar{\mathbf{x}}) = 1 \\ \infty & \text{if } \max_{0\leq n\leq t} \imath_k(\mathbf{X}; Y_k^n) \leq \lambda \\ & \quad \text{and } A_k^{(U)}(\bar{\mathbf{x}}) = 0 \end{cases} \tag{42}$$

where $\{A_k^{(u)}(\bar{\mathbf{x}})\}$ are independent Bernoulli distributed RVs with parameters $\max_{0\leq n\leq t} \mathbb{P}^{(u)}\Big[\imath_k(\mathbf{X}; Y_k^n) < \lambda | \mathbf{X} = \bar{\mathbf{x}}\Big] / \varepsilon_k^{(u)}(\bar{\mathbf{x}})$. The stopping time in (42) roughly states that if the information density of decoder k exceeds λ before time t, decoder k should send a stop signal when this happens. If this does not happen, the decoder should choose randomly between sending a stop signal at time t or letting $\hat{\tau}_k = \infty$. The key observation is that by replacing the stopping times τ_k by $\hat{\tau}_k$ in (36), the probability in (36) equals (40). We note that τ_k cannot be chosen equal to $\hat{\tau}_k$ in an achievability scheme because τ_k is only defined with respect to the filtration $\{\sigma(U, Y_k^n)\}_n$, whereas $\hat{\tau}_k$ is defined with respect to the larger filtration $\{\sigma(U, X^n, Y_k^n)\}_n$. This is, however, not issue in the converse argument. Note also that the auxiliary stopping times $\{\hat{\tau}_k\}$ are different for each t.

Next, by substituting (41) in (33), we conclude that

$$\mathbb{E}\Big[\max_k \tau_k | U = u, \mathbf{X} = \bar{\mathbf{x}}\Big] \geq \sum_{t=0}^{\infty}\Big(1 - L_t(\varepsilon_1^{(u)}(\bar{\mathbf{x}}), \cdots, \varepsilon_K^{(u)}(\bar{\mathbf{x}}))\Big). \tag{43}$$

Hence, $\mathbb{E}[\max_k \tau_k]$ can be lower-bounded as follows:

$$\mathbb{E}\Big[\max_k \tau_k\Big] \geq \min_{\substack{P_{U,\mathbf{X}} \in \mathcal{P}(\mathcal{U}\times\mathcal{X}^\infty), \varepsilon_k^{(u)}(\mathbf{x}) \in [0,1]: \\ \mathbb{E}\left[\varepsilon_k^{(U)}(\mathbf{X})\right] \leq \epsilon+\eta}} \mathbb{E}\left[\sum_{t=0}^{\infty}\Big(1 - L_t(\varepsilon_1^{(U)}(\mathbf{X}), \cdots, \varepsilon_K^{(U)}(\mathbf{X}))\Big)\right]. \tag{44}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\prod_k \mathbb{P}^{(u)}\Big[\mathsf{len}\big(\bar{\mathbf{Y}}_k\big) \le t \Big| \mathbf{X}=\bar{\mathbf{x}}\Big]$$

$$= \prod_k \Bigg(\mathbb{P}^{(u)}\Bigg[\max_{0\le n\le \mathsf{len}\left(\bar{\mathbf{Y}}_k\right)} i_k(\mathbf{X};\bar{Y}_k^n) \ge \lambda, \mathsf{len}\big(\bar{\mathbf{Y}}_k\big) \le t \Big| \mathbf{X}=\bar{\mathbf{x}}\Bigg] + \mathbb{P}^{(u)}\Bigg[\max_{0\le n\le \mathsf{len}\left(\bar{\mathbf{Y}}_k\right)} i_k(\mathbf{X};\bar{Y}_k^n) < \lambda, \mathsf{len}\big(\bar{\mathbf{Y}}_k\big) \le t \Big| \mathbf{X}=\bar{\mathbf{x}}\Bigg]\Bigg) \tag{36}$$

$$\le \prod_k \min\Bigg\{1, \mathbb{P}^{(u)}\Bigg[\max_{0\le n\le \min\left\{t,\mathsf{len}\left(\bar{\mathbf{Y}}_k\right)\right\}} i_k(\mathbf{X};\bar{Y}_k^n) \ge \lambda \Big| \mathbf{X}=\bar{\mathbf{x}}\Bigg] + \mathbb{P}^{(u)}\Bigg[\max_{0\le n\le \mathsf{len}\left(\bar{\mathbf{Y}}_k\right)} i_k(\mathbf{X};\bar{Y}_k^n) < \lambda \Big| \mathbf{X}=\bar{\mathbf{x}}\Bigg]\Bigg\} \tag{37}$$

$$\le \prod_k \min\Bigg\{1, \mathbb{P}^{(u)}\Bigg[\max_{0\le n\le \min\left\{t,\mathsf{len}\left(\bar{\mathbf{Y}}_k\right)\right\}} i_k(\mathbf{X};\bar{Y}_k^n) \ge \lambda \Big| \mathbf{X}=\bar{\mathbf{x}}\Bigg] + \mathbb{P}^{(u)}\Big[i_k(\mathbf{X};\bar{\mathbf{Y}}_k) < \lambda \Big| \mathbf{X}=\bar{\mathbf{x}}\Big]\Bigg\} \tag{38}$$

$$= \prod_k \min\Bigg\{1, \mathbb{P}^{(u)}\Bigg[\max_{0\le n\le \min\left\{t,\mathsf{len}\left(\bar{\mathbf{Y}}_k\right)\right\}} i_k(\mathbf{X};\bar{Y}_k^n) \ge \lambda \Big| \mathbf{X}=\bar{\mathbf{x}}\Bigg] + \varepsilon_k^{(u)}(\bar{\mathbf{x}})\Bigg\} \tag{39}$$

$$\le \max_{x^t\in\mathcal{X}^t} \prod_k \min\Big\{1, \mathbb{P}\Big[\max_{0\le n\le t} i_k(x^n;Y_k^n) \ge \lambda\Big] + \varepsilon_k^{(u)}(\bar{\mathbf{x}})\Big\} \tag{40}$$

$$= L_t(\varepsilon_1^{(u)}(\bar{\mathbf{x}}), \cdots, \varepsilon_K^{(u)}(\bar{\mathbf{x}})). \tag{41}$$

The right-hand side of (44) depends on the code only through the random quantities $\varepsilon_k^{(U)}(\mathbf{X})$. Now, by defining $\overline{U}=(U,\mathbf{X})$ and $\overline{\mathcal{U}}=\mathcal{U}\times\mathcal{X}^\infty$, we obtain

$$\mathbb{E}\Big[\max_k \tau_k\Big] \ge \min_{\substack{P_{\overline{U}}\in\mathcal{P}(\overline{\mathcal{U}}),\varepsilon_k^{(u)}\in[0,1]:\\ \mathbb{E}\left[\varepsilon_k^{(\overline{U})}\right]\le\epsilon+\eta}} \mathbb{E}\left[\sum_{t=0}^{\infty}\Big(1-L_t(\varepsilon_1^{(\overline{U})},\cdots,\varepsilon_K^{(\overline{U})})\Big)\right]. \tag{45}$$

This concludes the proof provided that one shows that the cardinality of $\overline{\mathcal{U}}$ in (45) can be upper-bounded by $K+1$. This cardinality bound, which can be established by an application of Caratheodory's theorem, is provided in Appendix III ∎

We remark that the converse bound in Theorem 3 also provides a new converse for the single-decoder setup when $K=1$. However, when evaluated numerically, it turns out that this bound is less tight than the converse bounds for full feedback provided in [8].

As we shall see next, choosing $Q_k^{(\infty)}$ as a simple product distribution yields a computable and tight nonasymptotic bound for symmetric channels.[5] For general CM-DMBCs, choosing $Q_k^{(\infty)}$ as a convex combination of product distributions (cf., (159)) appears necessary to obtain tight large-ℓ asymptotic expansions.

If $\{W_k\}$ are identical and symmetric, we have the following particularization of Theorem 3.

Corollary 4: For arbitrary $M\in\mathbb{N}$, $t\in\mathbb{Z}_+$, $\eta>0$, and an arbitrary sequence $\mathbf{x}\in\mathcal{X}^\infty$, let

$$v_t = \mathbb{P}\Big[\max_{0\le n\le t} i_{P^*,W_1}(x^n;Y_1^n) \ge \log M + \log\eta\Big] \tag{46}$$

where $Y_1^t\sim P_{Y_1^t|X^t=x^t}$ and $i_{P^*,W_1}(\cdot;\cdot)$ is defined in (8). When $W_1=\cdots=W_K$ and W_1 is symmetric, every (ℓ,M,ϵ)-VLSF code must satisfy

$$\ell \ge \min_{\substack{P_U\in\mathcal{P}(\mathcal{U}),\varepsilon^{(u)}\in[0,1]:\\ \mathbb{E}_U\left[\varepsilon^{(U)}\right]\le\epsilon+\eta}} \mathbb{E}_U\left[\sum_{t=0}^{\infty}\Big(1-\min\Big\{1,v_t+\varepsilon^{(U)}\Big\}^K\Big)\right]. \tag{47}$$

Proof: We apply the converse bound in Theorem 3 with $Q_1^{(\infty)}=\cdots=Q_K^{(\infty)}$. Furthermore, we choose $Q_1^{(\infty)}$ as a product distribution with marginal $Q_1(y)=1/|\mathcal{Y}_1|$ for all $y\in\mathcal{Y}_1$. By (22), we have that $Q_k(\overline{\mathbf{y}}_k)=|\mathcal{Y}_k|^{-\mathsf{len}(\overline{\mathbf{y}}_k)}$. Since the capacity-achieving output distribution of a symmetric channel is uniform [17, Eq. (7.22)], we conclude that

$$i_k(x^t;Y_k^t) \sim i_{P^*,W_1}(x^t,Y_1^t) \tag{48}$$

for all $k\in\mathcal{K}$ given that $X^t=x^t$. One can verify that, when the channel is symmetric, the conditional probability distribution of the information density $i_{P^*,W_1}(x^t,Y_1^t)$ given $X^t=x^t$ does not depend on x^t. This allows us to drop the maximization over x^t in (29). Finally, to express the minimization problem (30) in the form given in (47), we note that, for every $[\varepsilon_1,\cdots,\varepsilon_K]\in[0,1]^K$, we have

$$\Big(\prod_{k=1}^K \min\{1,v_t+\varepsilon_k\}\Big)^{1/K} \le \frac{1}{K}\sum_{k=1}^K \min\{1,v_t+\varepsilon_k\} \tag{49}$$

$$\le \min\Big\{1, v_t+\frac{1}{K}\sum_{k=1}^K\varepsilon_k\Big\}. \tag{50}$$

Here, (49) follows because the geometric mean is no larger than the arithmetic mean and (50) follows from Jensen's inequality. ∎

[5] A channel is symmetric if the rows and columns of the stochastic channel matrix are permutations of each other [17, p. 189].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

C. Asymptotic Expansion

Analyzing (19), (21) and (30) in the limit $\ell \to \infty$, we obtain the following asymptotic characterization of $\log M^*_{\text{sf}}(\ell, \epsilon)$.

Theorem 5: Let $V \triangleq (\prod_k V_k)^{1/K}$ and $\varrho_k \triangleq \sqrt{V_k/V}$ and assume, without loss of generality, that $C_1 \geq \cdots \geq C_K$. For every CM-DMBC satisfying

$$\frac{1}{C_i} + \frac{i}{C_K} > \frac{i}{C}, \qquad i \in \{1, \cdots, K-1\} \tag{51}$$

and for every $\epsilon \in (0,1)$, we have[6]

$$\begin{aligned} &\frac{C\ell}{1-\epsilon} - \sqrt{\frac{V\ell}{1-\epsilon}}\Xi_{\text{a}} + o(\sqrt{\ell}) \\ &\leq \log M^*_{\text{sf}}(\ell, \epsilon) \leq \frac{C\ell}{1-\epsilon} - \sqrt{\frac{V\ell}{1-\epsilon}}\Xi_{\text{c}} + o(\sqrt{\ell}). \end{aligned} \tag{52}$$

Here,

$$\Xi_{\text{a}} \triangleq \min_{\mathbf{v} \in \mathbb{R}_0^{|\mathcal{X}|}} \mathbb{E}\Big[\max_k \nabla I_k(\mathbf{v}) + \varrho_k Z_k\Big] \tag{53}$$

and

$$\Xi_{\text{c}} \triangleq \mathbb{E}\Big[\max_k H_k\Big] \tag{54}$$

where $Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\{H_k\}$ are independent RVs with cumulative distribution functions

$$F_{H_k}(w) \triangleq \Phi\left(\frac{w + \nabla I_k(\hat{\mathbf{v}}(w))}{\varrho_k}\right). \tag{55}$$

The function $\hat{\mathbf{v}}(\cdot)$ is defined as follows:[7]

$$\hat{\mathbf{v}}(w) \triangleq \arg\max_{\mathbf{v} \in \mathbb{R}_0^{|\mathcal{X}|}} \prod_k \Phi\left(\frac{w + \nabla I_k(\mathbf{v})}{\varrho_k}\right). \tag{56}$$

Proof: The converse bound in (52) is proved in Appendix IV and the achievability bound in (52) is proved in Appendix V. We next provide a heuristic argument that sheds light on the achievability part. The key step is to obtain a tight upper bound on $\mathbb{E}\Big[\max_k \tau_k^{(0)}\Big]$ in (19). The desired asymptotic expansion then follows by properly choosing γ and q in Theorem 1 (see Appendix V for details). For the purpose of this heuristic argument, consider the special case $C_1 = \ldots = C_K = C$, and $V_1 = \ldots = V_K$. Also let us assume that, when $P_{X^\infty} = (P^*)^\infty$, the information densities can be well-approximated by the Brownian motions with drift

$$\imath_k(X^n; Y^n) \approx nC + \sqrt{V_1}B(n) \tag{57}$$

where $B(n)$ is a standard Brownian motion. It now follows from the Bachelier-Levy formula (see [18] or [19]) that the probability density function of the first passage time $\inf\{t \in \mathbb{R}_+ : tC + \sqrt{V_1}B(t) \geq \gamma\}$ is given by

$$\frac{\gamma}{\sqrt{V_1}t^{3/2}}\phi\left(\frac{\gamma - tC}{\sqrt{tV_1}}\right) \tag{58}$$

where $\phi(\cdot)$ is the probability density function for the standard Gaussian RV. This shows that

$$\frac{\tau_k^{(0)} - \gamma/C}{\sqrt{\gamma V_1/C^3}} \xrightarrow{\text{d}} \mathcal{N}(0,1) \tag{59}$$

as $\gamma \to \infty$ and, as a consequence, we have that

$$\mathbb{E}\left[\max_k \frac{\tau_k^{(0)} - \gamma/C}{\sqrt{\gamma V_1/C^3}}\right] \to \mathbb{E}\Big[\max_k Z_k\Big]. \tag{60}$$

Rewriting (60), we obtain

$$\mathbb{E}\Big[\max_k \tau_k^{(0)}\Big] = \frac{\gamma}{C} + \sqrt{\frac{\gamma V_1}{C^3}}\mathbb{E}\Big[\max_k Z_k\Big] + o(\sqrt{\gamma}). \tag{61}$$

This result is a particularization of Lemma 12 in Appendix V for the case where $C_1 = \ldots = C_K = C$ and where $V_1 = \ldots = V_K$. Next, let δ be an arbitrary positive constant. By choosing $\gamma = \frac{\ell C}{1-\epsilon} - (1+\delta)\sqrt{\frac{V_1\ell}{1-\epsilon}}\mathbb{E}\Big[\max_k Z_k\Big]$ and by setting $q = \epsilon - \Theta(1/\ell)$, we observe from (61) that $(1-q)\mathbb{E}\Big[\max_k \tau_k^{(0)}\Big] \leq \ell$ for all sufficiently large ℓ. As a result, Theorem 1 implies the following asymptotic expansion of $\log M^*_{\text{sf}}(\ell, \epsilon)$

$$\begin{aligned} &\log M^*_{\text{sf}}(\ell, \epsilon) \\ &\quad \geq \gamma + \log\frac{\epsilon - q}{1-q} && (62) \\ &\quad \geq \frac{\ell C}{1-\epsilon} - (1+\delta)\sqrt{\frac{V_1\ell}{1-\epsilon}}\mathbb{E}\Big[\max_k Z_k\Big] + o(\sqrt{\ell}) && (63) \end{aligned}$$

This argument is made rigorous and further generalized in Appendix V. ∎

Some remarks are in order. The condition (51) is needed only for the converse part. Furthermore, when $K = 2$, the condition (51) reduces to $1/C_1 + 1/C_2 > 1/C$. Note that for every CM-DMBC, we have that $1/C_1 + 1/C_2 \geq 1/C$. Indeed, suppose on the contrary that $1/C_1 + 1/C_2 < 1/C$. Then one can achieve a rate larger than C by sequential transmission to the two decoders:

$$\begin{aligned} &\max_{\alpha \in [0,1]} \min\{\alpha C_1, (1-\alpha)C_2\} \\ &\quad = \min\left\{\frac{C_2}{C_1 + C_2}C_1, \left(1 - \frac{C_2}{C_1 + C_2}\right)C_2\right\} && (64) \\ &\quad = \frac{1}{1/C_1 + 1/C_2} && (65) \\ &\quad > C. && (66) \end{aligned}$$

But this contradicts the fact that C is the capacity. Theorem 5 does not hold for the special case $1/C_1 + 1/C_2 = 1/C$.

As we shall show next, the constants Ξ_{a} and Ξ_{c} defined in (53) and (54), respectively, satisfy $\Xi_{\text{a}} \geq \Xi_{\text{c}} > 0$. This implies that the second-order term in the asymptotic expansion of $\log M^*_{\text{sf}}(\ell, \epsilon)$ is positive for every $\epsilon \in (0,1)$ (see (52))—a result that strengthens [13, Th. 3]. Proving that $\Xi_{\text{a}} \geq \Xi_{\text{c}}$ will also allow us to shed light on the reason behind the gap between the achievability and converse bound and the role of the RVs $\{H_k\}$ in (54).

Proposition 6: Under the conditions described in Theorem 5, the constants in (53) and (54) satisfy

$$0 < \Xi_{\text{a}} \leq \Xi_{\text{c}}. \tag{67}$$

[6]The subscripts "a" and "c" in Ξ_{a} and Ξ_{c} stand for achievability and converse, respectively.

[7]If the maximizer in (56) is not unique, $\hat{\mathbf{v}}(w)$ is chosen arbitrarily from the set of maximizers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Proof: It is convenient to rewrite Ξ_{a} as follows:

$$\Xi_{\mathrm{a}} = \min_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \lim_{n\to\infty} \mathbb{E}\Big[\max_k (\nabla I_k(\mathbf{v}) + n + \varrho_k Z_k)^+ - n\Big]. \quad (68)$$

To obtain (68), we used that for every set $\{X_k\}$ of integrable RVs, $\max_k (X_k+n)^+ - n$ converges in distribution to $\max_k X_k$ as $n\to\infty$. Hence, by using that $|\max_k (X_k+n)^+ - n| \le \max_k |\max\{-n, X_k\}| \le \max|X_k|$ almost surely for all $n\in\mathbb{N}$, we invoke Lebesgue's dominated convergence theorem [14, Th. 16.4] to conclude that

$$\lim_{n\to\infty} \mathbb{E}\Big[\max_k (X_k+n)^+ - n\Big] = \mathbb{E}\Big[\max_k X_k\Big]. \quad (69)$$

This implies (68). Next, we bound (68) as follows

$$\begin{aligned}
&\Xi_{\mathrm{a}} \\
&= \min_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \lim_{n\to\infty} \int_0^\infty \left(1 - \prod_k \Phi\left(\frac{w-n-\nabla I_k(\mathbf{v})}{\varrho_k}\right)\right)\mathrm{d}w \\
&\qquad - n \quad (70)\\
&\ge \lim_{n\to\infty} \int_0^\infty \left(1 - \prod_k \Phi\left(\frac{w-n+\nabla I_k(\hat{\mathbf{v}}(w-n))}{\varrho_k}\right)\right)\mathrm{d}w \\
&\qquad - n \quad (71)\\
&= \lim_{n\to\infty} \mathbb{E}\Big[\max_k (H_k+n)^+ - n\Big] \quad (72)\\
&= \mathbb{E}\Big[\max_k H_k\Big] \quad (73)\\
&= \Xi_{\mathrm{c}}. \quad (74)
\end{aligned}$$

To obtain (70), we used that for every nonnegative RV X, we have $\int_0^\infty (1-\mathbb{P}[X<x])\mathrm{d}x = \mathbb{E}[X]$; we also used that $\{Z_k\}$ are i.i.d. Gaussian and that for every real-valued RV T and every $w\ge 0$, we have that $\mathbb{P}[(T)^+ < w] = \mathbb{P}[T<w]$. The inequality (71) follows from (56). Finally, (72) and (73) follow from steps similar to the ones leading to (70) and (68), respectively.

The inequality (71) reveals the origin of the gap between Ξ_{a} and Ξ_{c}. The constant Ξ_{a} is obtained by evaluating the achievability bound in Theorem 1 for an i.i.d. process X^∞. Instead, in the computation of Ξ_{c}, we find the input distribution that maximizes $\mathbb{P}[\max_k \tau_k \le t]$ for each t. The resulting process is not i.i.d.

To prove that $\Xi_{\mathrm{c}} > 0$, we proceed as follows:

$$\begin{aligned}
&\Xi_{\mathrm{c}} \\
&= \lim_{n\to\infty} \int_0^\infty \left(1 - \prod_k \Phi\left(\frac{w-n+\nabla I_k(\hat{\mathbf{v}}(w-n))}{\varrho_k}\right)\right)\mathrm{d}w \\
&\qquad - n \quad (75)\\
&> \lim_{n\to\infty} \int_0^\infty \left(1 - \min_k \Phi\left(\frac{w-n+\nabla I_k(\hat{\mathbf{v}}(w-n))}{\varrho_k}\right)\right)\mathrm{d}w \\
&\qquad - n \quad (76)\\
&\ge \lim_{n\to\infty} \int_0^\infty \left(1 - \min_k \Phi\left(\frac{w-n}{\varrho_k}\right)\right)\mathrm{d}w - n \quad (77)\\
&\ge \lim_{n\to\infty} \min_k \int_0^\infty \left(1 - \Phi\left(\frac{w-n}{\varrho_k}\right)\right)\mathrm{d}w - n \quad (78)\\
&= \lim_{n\to\infty} \min_k \mathbb{E}\big[(\varrho_k Z_k + n)^+ - n\big] = \min_k \mathbb{E}[\varrho_k Z_k] = 0. \quad (79)
\end{aligned}$$

Here, (75) follows from (71); the inequality in (76) holds because $\prod_k a_k < \min_k a_k$ for all $a_k \in (0,1)$; finally, (77) follows because $\min_k \nabla I_k(\hat{\mathbf{v}}(w)) \le 0$ for all $w\in\mathbb{R}$. Indeed, $\nabla I_k(\cdot)$ is the directional derivative of $I_k(P)$ computed at the unique capacity-achieving input distribution P^*. The last equation (79) follows from an argument similar to the one leading to (69). ■

There are cases where $\Xi_{\mathrm{a}} = \Xi_{\mathrm{c}}$, and hence (52) provides a complete second-order characterization of $\log M^*_{\mathrm{sf}}(\ell,\epsilon)$. This happens when $\hat{\mathbf{v}}(\cdot)$ in (56) equals $\mathbf{0}_{|\mathcal{X}|}$, which occurs for example when P^* simultaneously maximizes $I_k(P)$ for all $k\in\mathcal{K}$. In the following corollary, we provide sufficient conditions for Theorem 5 to yield an asymptotic expansion that is tight up to the second order.

Corollary 7: We have that $\Xi_{\mathrm{a}} = \Xi_{\mathrm{c}}$ and, hence,

$$\log M^*_{\mathrm{sf}}(\ell,\epsilon) = \frac{C\ell}{1-\epsilon} - \sqrt{\frac{V\ell}{1-\epsilon}}\,\mathbb{E}\Big[\max_k Z_k\Big] + o(\sqrt{\ell}) \quad (80)$$

if either of the following conditions hold

1) The capacity-achieving input distribution P^* simultaneously maximizes $I_k(P)$ for all $k\in\mathcal{K}$, or
2) $V_1 = \cdots = V_K$ and

$$\sum_k \nabla I_k(\mathbf{e}_{|\mathcal{X}|}(x)) = 0, \qquad x\in\{1,\cdots,|\mathcal{X}|\}. \quad (81)$$

Here, $\mathbf{e}_{|\mathcal{X}|}(i)$ denotes the $|\mathcal{X}|$-dimensional vector whose ith entry is equal to one and whose remaining entries are equal to zero.

Proof: We shall prove that $\Xi_{\mathrm{a}} = \Xi_{\mathrm{c}}$ under the stated conditions by characterizing $\hat{\mathbf{v}}(\cdot)$ in (56). When the component channels $W_1,\cdots,W_K$ have the same capacity-achieving input distribution, we have $\nabla I_k(\mathbf{v}) = 0$ for all $\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}$ and all $k\in\mathcal{K}$. Hence, $\mathbf{0}_{|\mathcal{X}|}$ is a maximizer of (56) which implies that (71) holds with equality.

Consider now the case that $\nabla I_k(\mathbf{v}) \neq 0$ for some $\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}$ and some $k\in\mathcal{K}$. Let $\ker(\nabla I_k)$ denote the kernel of ∇I_k, and let $\mathcal{D}^\dagger = \mathbb{R}_0^{|\mathcal{X}|} \cap \bigcap_{k=1}^K \ker(\nabla I_k)$ and $\mathcal{D} = \mathcal{R}_0^{|\mathcal{X}|} \setminus \mathcal{D}^\dagger$. We note that $\mathcal{D}\neq\emptyset$ since, by assumption, there exists a $\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}$ and a $k\in\mathcal{K}$ such that $\nabla I_k(\mathbf{v})\neq 0$. Let the dimension of the linear subspace $\mathcal{D}$ be m and let $\mathbb{D}\in\mathbb{R}^{|\mathcal{X}|\times m}$ be an $|\mathcal{X}|$-by-m matrix with columns spanning the linear subspace $\mathcal{D}$. Now, define

$$\hat{\mathbf{v}}_{\mathcal{D}}(w) \triangleq \operatorname*{arg\,max}_{\mathbf{v}_{\mathcal{D}}\in\mathbb{R}^m} \prod_k \Phi\left(\frac{w+\nabla I_k(\mathbb{D}\mathbf{v}_{\mathcal{D}})}{\varrho_k}\right) \quad (82)$$

and

$$\tilde{\mathbf{v}}_{\mathcal{D}} \triangleq \operatorname*{arg\,min}_{\mathbf{v}_{\mathcal{D}}\in\mathbb{R}^m} \lim_{n\to\infty} \left(\int_0^\infty \left(1 - \prod_k \Phi\left(\frac{w-n-\nabla I_k(\mathbb{D}\mathbf{v}_{\mathcal{D}})}{\varrho_k}\right)\right)\mathrm{d}w - n\right). \quad (83)$$

We note that $\hat{\mathbf{v}}_{\mathcal{D}}(w)$ is continuous in w and that log-concavity of the objective function in (82) implies that $\hat{\mathbf{v}}_{\mathcal{D}}(w)$ is the unique maximizer of the optimization problem. Moreover, $\mathbb{D}\hat{\mathbf{v}}_{\mathcal{D}}(w)$ is a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

maximizer of (56) and $\mathbb{D}\tilde{\mathbf{v}}_{\mathcal{D}}$ is a minimizer in (70). This implies that the steps (70)–(74) can be equivalently written as

$$\begin{aligned}
&\Xi_{\mathrm{a}} \\
&= \min_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \lim_{n\to\infty} \\
&\quad \int_0^\infty \left(1-\prod_k \Phi\left(\frac{w-n-\nabla I_k(\mathbf{v})}{\varrho_k}\right)\right)\mathrm{d}w - n && (84)\\
&= \lim_{n\to\infty} \\
&\quad \int_0^\infty \left(1-\prod_k \Phi\left(\frac{w-n-\nabla I_k(\mathbb{D}\tilde{\mathbf{v}}_{\mathcal{D}})}{\varrho_k}\right)\right)\mathrm{d}w - n && (85)\\
&\geq \lim_{n\to\infty} \int_0^\infty \Bigg(1 \\
&\qquad -\prod_k \Phi\left(\frac{w-n+\nabla I_k(\mathbb{D}\hat{\mathbf{v}}_{\mathcal{D}}(w-n))}{\varrho_k}\right)\Bigg)\mathrm{d}w - n && (86)\\
&= \lim_{n\to\infty} \int_0^\infty \Bigg(1 \\
&\qquad -\prod_k \Phi\left(\frac{w-n+\nabla I_k(\hat{\mathbf{v}}(w-n))}{\varrho_k}\right)\Bigg)\mathrm{d}w - n && (87)\\
&= \Xi_{\mathrm{c}}. && (88)
\end{aligned}$$

Since $\hat{\mathbf{v}}_{\mathcal{D}}(w)$ is continuous and is the unique maximizer of (82), the inequality in (86) holds with equality if and only if $\hat{\mathbf{v}}_{\mathcal{D}}(w) = \mathbf{a}$ almost everywhere for some vector $\mathbf{a} \in \mathbb{R}^m$ that does not depend on w.

Suppose that $\hat{\mathbf{v}}_{\mathcal{D}}(w) = \mathbf{a}$. The objective function in (82) is positive, strictly log-concave in $\mathbf{v}_{\mathcal{D}} \in \mathbb{R}^m$, and tends to zero as $\|\mathbf{v}_{\mathcal{D}}\| \to \infty$. Thus, the unique maximum is at the unique stationary point, which can be found by differentiating the logarithm of the objective function in (82) in each of the m dimensions and by equating it to zero. This yields

$$\sum_k \psi\left(\frac{w+\nabla I_k(\mathbb{D}\mathbf{a})}{\varrho_k}\right)\frac{\nabla I_k(\mathbb{D}\mathbf{e}_m(i))}{\varrho_k} = 0, \quad i\in\{1,\cdots,m\}, w\in\mathbb{R} \qquad (89)$$

where $\psi(w) \triangleq \phi(w)/\Phi(w)$. It follows that (89) cannot be satisfied for every $w \in \mathbb{R}$ unless $\mathbf{a} = \mathbf{0}_m$ and $\varrho_1 = \cdots = \varrho_K$. In this case (89) reduces to (81). ∎

For broadcast channels that do not satisfy the conditions of Corollary 7, we can tighten the left-hand side of (52) by using an input distribution that is not stationary memoryless. This yields the following theorem.

Theorem 8: Let $V \triangleq (\prod_k V_k)^{1/K}$ and $\varrho_k \triangleq \sqrt{V_k/V}$. Fix a differentiable function $\bar{\mathbf{v}} : \mathbb{R} \mapsto \mathbb{R}_0^{|\mathcal{X}|}$ such that

$$P^* + C\bar{\mathbf{v}}'(w) \in \mathcal{P}(\mathcal{X}) \qquad (90)$$

for all $w \in \mathbb{R}$. Additionally, define

$$E_k(s) \triangleq C - I_k(P^* + C\bar{\mathbf{v}}'(s)) + C\nabla I_k(\bar{\mathbf{v}}'(s)) \qquad (91)$$

and assume that

$$\int_{-\infty}^{\infty} E_k(s)\mathrm{d}s < \infty \qquad (92)$$

and that

$$\sup_s |E_k'(s)| < \infty. \qquad (93)$$

Then, for every CM-DMBC, we have

$$\log M_{\mathrm{sf}}^*(\ell,\epsilon) \geq \frac{C\ell}{1-\epsilon} - \sqrt{\frac{V\ell}{1-\epsilon}}\bar{\Xi}_{\mathrm{a}} - o(\sqrt{\ell}). \qquad (94)$$

Here,

$$\bar{\Xi}_{\mathrm{a}} \triangleq \mathbb{E}\left[\max_k \bar{H}_k\right] \qquad (95)$$

where the independent RVs $\{\bar{H}_k\}$ are defined by the cumulative distribution functions

$$F_{\bar{H}_k}(w) \triangleq \Phi\left(\frac{1}{\varrho_k}\left(w+\nabla I_k(\bar{\mathbf{v}}(w)) - \int_{-\infty}^{w}\frac{E_k(s)}{C}\mathrm{d}s\right)\right). \qquad (96)$$

Proof: See Appendix V. ∎

If one sets $\bar{\mathbf{v}}(\cdot)$ in Theorem 8 equal to $\hat{\mathbf{v}}(\cdot)$ in (56), the resulting gap between Ξ_{c} and $\bar{\Xi}_a$ is caused only by the "error" term $E_k(s)$. Interestingly, there are channels beyond the ones for which Corollary 7 applies where $E_k(s) = 0$ and, hence, a complete second-order characterization of $\log M_{\mathrm{sf}}^*(\ell,\epsilon)$ is available. The next corollary describes a class of channels for which this is the case.

Corollary 9: Let $\mathcal{X}_1,\cdots,\mathcal{X}_R$ be disjoint sets and let $\mathcal{X} = \cup_{r=1}^R \mathcal{X}_r$. Moreover, for $k \in \mathcal{K}$ and $r \in \{1,\cdots,R\}$, let $W_{k,r}$ be a channel from $\mathcal{X}_r$ to $\mathcal{Y}_k$ with capacity-achieving input distribution P_r^* (independent of k), capacity-achieving output distribution $P_{Y_k}^*$ (independent of r), and capacity $C_{k,r}$. Define for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}_k$ the channel $W_k(y|x) = W_{k,r(x)}(y|x)$, where the function $r : \mathcal{X} \mapsto \{1,\cdots,R\}$ is such that $x \in \mathcal{X}_{r(x)}$. Assume that $C_1 \geq \cdots \geq C_K$ and that

$$\frac{1}{C_i} + \frac{i}{C_K} > \frac{i}{C}, \qquad i \in \{1,\cdots,K-1\}. \qquad (97)$$

Define

$$\boldsymbol{\beta}(w) \triangleq \arg\max_{\boldsymbol{\beta}\in\mathbb{R}_0^R} \prod_k \Phi\left(\frac{1}{\varrho_k}\left(w+\sum_{r=1}^R \beta_r C_{k,r}\right)\right) \qquad (98)$$

and assume that

$$P^*(x) + CP_{r(x)}^*(x)\beta_{r(x)}'(w) \in [0,1] \qquad (99)$$

for every $x \in \mathcal{X}$ and $w \in \mathbb{R}$. Then, for every $\epsilon \in (0,1)$,

$$\log M_{\mathrm{sf}}^*(\ell,\epsilon) = \frac{C\ell}{1-\epsilon} - \sqrt{\frac{V\ell}{1-\epsilon}}\Xi_{\mathrm{c}} + o(\sqrt{\ell}) \qquad (100)$$

where Ξ_{c} is defined in (54).

Proof: We shall first evaluate the mutual information and the directional derivative of the mutual information. Define the input distribution

$$P_{\boldsymbol{\alpha}} \triangleq [\alpha_1 P_1^*, \alpha_2 P_2^*, \cdots, \alpha_R P_R^*] \qquad (101)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

for all nonnegative vectors $\boldsymbol{\alpha}$ with $\sum_{r=1}^{R}\alpha_r = 1$. The mutual information $I_k(P_{\boldsymbol{\alpha}})$ is given by

$$\begin{aligned} &I_k(P_{\boldsymbol{\alpha}}) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}_k} P_{\boldsymbol{\alpha}}(x)W_k(y|x)\log\frac{W_k(y|x)}{P^*_{Y_k}(y)} && (102)\\ &= \sum_{r=1}^{R}\alpha_r\sum_{x\in\mathcal{X}_r}\sum_{y\in\mathcal{Y}_k} P^*_r(x)W_{k,r}(y|x)\log\frac{W_{k,r}(y|x)}{P^*_{Y_k}(y)} && (103)\\ &= \sum_{r=1}^{R}\alpha_r C_{k,r}. && (104)\end{aligned}$$

In (104), we used that the channels $W_{k,r}$ have the same capacity-achieving output distribution for $r \in \{1,\cdots,R\}$.

Next, we let $\boldsymbol{\alpha}^*$ be the maximizer of $\boldsymbol{\alpha}\mapsto\min_k I_k(P_{\boldsymbol{\alpha}})$ and compute the directional derivative of the mutual information at $P_{\boldsymbol{\alpha}^*}$ along the direction $\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}$:

$$\begin{aligned}\nabla_{P_{\boldsymbol{\alpha}^*}} I_k(\mathbf{v}) &= \sum_{r=1}^{R}\sum_{x\in\mathcal{X}_r} v_x D(W_{k,r}(\cdot|x)||P^*_{Y_k}) && (105)\\ &= \sum_{r=1}^{R}\left(\sum_{x\in\mathcal{X}_r} v_x\right)C_{k,r}. && (106)\end{aligned}$$

Here, (105) follows because the output distribution at decoder k given an input distribution of the form (101) is $P^*_{Y_k}$ and (106) follows from the assumption $P^*(x)>0$, which implies that $P^*_r(x) > 0$ for $x\in\mathcal{X}_r$ and $r\in\{1,\cdots,R\}$. In turn, this implies that $D(W_{k,r}(\cdot|x)||P^*_{Y_k}) = C_{k,r}$ (see, e.g., [20, Th. 4.5.1]). It follows from (104) and (106) that the capacity C is achieved using time-sharing and that the capacity-achieving input distribution P^* must have the form given by (101). Indeed, by the concavity of mutual information and by the definition of $\boldsymbol{\alpha}^*$, we have that, for all $P\in\mathcal{P}(\mathcal{X})$,

$$\begin{aligned}&\min_k I_k(P) \\ &\le \min_k\left\{I_k(P_{\boldsymbol{\alpha}^*}) + \nabla_{P_{\boldsymbol{\alpha}^*}} I_k(P-P_{\boldsymbol{\alpha}^*})\right\} && (107)\\ &= \min_k\left\{\sum_{r=1}^{R}\alpha^*_r C_{k,r} \right.\\ &\qquad \left. + \sum_{r=1}^{R}\left(\sum_{x\in\mathcal{X}_r}(P(x)-P_{\boldsymbol{\alpha}^*}(x))\right)C_{k,r}\right\} && (108)\\ &= \min_k\left\{\sum_{r=1}^{R}\left(\alpha^*_r + \left(\sum_{x\in\mathcal{X}_r}(P(x)-P_{\boldsymbol{\alpha}^*}(x))\right)\right)C_{k,r}\right\} && (109)\\ &\le \min_k I_k(P_{\boldsymbol{\alpha}^*}). && (110)\end{aligned}$$

Thus, we must have that $P_{\boldsymbol{\alpha}^*} = P^*$.

By substituting (106) in (55), we obtain

$$F_{H_k}(w) \triangleq \Phi\left(\frac{1}{\varrho_k}\left(w+\sum_{r=1}^{R}\beta_r(w)C_{k,r}\right)\right) \qquad (111)$$

where the function $\boldsymbol{\beta}:\mathbb{R}\mapsto\mathbb{R}_0^R$ is given by (98).

Next, we shall prove that we can achieve (100) using Theorem 8. Define the function $\bar{\mathbf{v}}:\mathbb{R}\mapsto\mathbb{R}^{|\mathcal{X}|}$ as follows:

$$\bar{v}_x(w) = P^*_{r(x)}(x)\beta_{r(x)}(w), \qquad x\in\mathcal{X}. \qquad (112)$$

Note that $\bar{\mathbf{v}}(w)$ maximizes (56). Indeed,

$$\begin{aligned}&\max_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}}\prod_k \Phi\left(\frac{w+\nabla I_k(\mathbf{v})}{\varrho_k}\right)\\ &= \max_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}}\prod_k \Phi\left(\frac{1}{\varrho_k}\left(w+\sum_{r=1}^{R}\left(\sum_{x\in\mathcal{X}_r} v_x\right)C_{k,r}\right)\right) && (113)\\ &= \max_{\boldsymbol{\beta}\in\mathbb{R}_0^R}\prod_k \Phi\left(\frac{1}{\varrho_k}\left(w+\sum_{r=1}^{R}\sum_{x\in\mathcal{X}_r} P^*_r(x)\beta_r C_{k,r}\right)\right) && (114)\\ &= \prod_k \Phi\left(\frac{1}{\varrho_k}\left(w+\sum_{x\in\mathcal{X}} P^*_{r(x)}(x)\beta_{r(x)}(w)C_{k,l}\right)\right) && (115)\\ &= \prod_k \Phi\left(\frac{1}{\varrho_k}\left(w+\nabla I_k(\bar{\mathbf{v}}(w))\right)\right). && (116)\end{aligned}$$

Here, (113) follows from (106) and (115) follows from the definition of $\boldsymbol{\beta}(w)$ in (98). Note that the definition of $\bar{\mathbf{v}}(w)$ implies that $P^*+C\bar{\mathbf{v}}'(w) = [(\alpha^*_1+C\beta'_1(w))P^*_1,\cdots,(\alpha^*_R+C\beta'_R(w))P^*_R]$, which is a probability distribution by the condition in (99) and has the form (101). Next, we have that

$$\begin{aligned}&E_k(s)\\ &= C - I_k(P^*+C\bar{\mathbf{v}}'(s)) + C\nabla I_k(\bar{\mathbf{v}}'(s)) && (117)\\ &= C - \sum_{r=1}^{R}(\alpha^*_r + C\beta'_r(w))C_{k,r} + C\sum_{r=1}^{R}\beta'_r(s)C_{k,r} && (118)\\ &= 0. && (119)\end{aligned}$$

Here, (118) follows from (104) and (106). Since $E_k(s)=0$, we have that H_k has the same distribution as $\bar{H}_k$ for $k\in\mathcal{K}$. Furthermore, $\int_{-\infty}^{\infty}E_k(s)\mathrm{d}s = 0$ and $|E'_k(s)|<\infty$ for every $s\in\mathbb{R}$. The conditions in Theorem 8 are thus satisfied, which implies that (100) is indeed achievable. ∎

The following lemma shows that there exist nontrivial channels that satisfy the conditions of Corollary 9.

Lemma 10: Let $R=2$, $K=2$, and $\Delta_1 \triangleq C_{11}-C_{12} > 0 > C_{21}-C_{22}\triangleq\Delta_2$. Let also

$$D \triangleq -\frac{\Delta_1\varrho_2^2+\Delta_2\varrho_1^2}{\Delta_1^2\varrho_2^2+\Delta_2^2\varrho_1^2}. \qquad (120)$$

Then, the condition $P^*(x)+CP^*_{r(x)}(x)\beta'_{r(x)}(w)\in[0,1]$ holds for every $x\in\mathcal{X}$ and every $w\in\mathbb{R}$ provided that

$$P^*(x)+(-1)^{r(x)+1}CP^*_{r(x)}(x)D\in[0,1] \qquad (121)$$

for every $x\in\mathcal{X}$ and

$$\left(\frac{\Delta_1}{\varrho_1}+\frac{\Delta_2}{\varrho_2}\right)(\varrho_2-\varrho_1)\ge 0. \qquad (122)$$

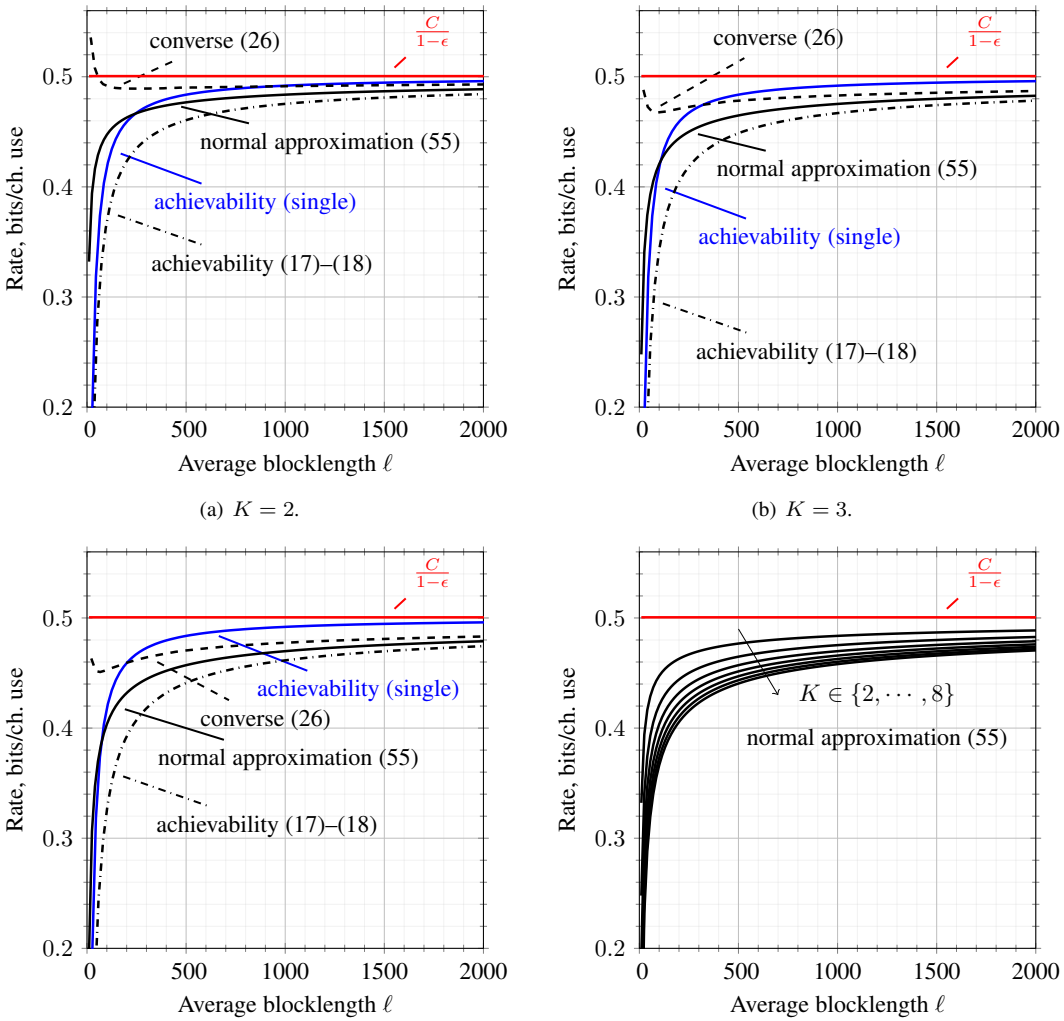
Proof: See Appendix VI. ∎

IV. Numerical Examples

A. Binary Symmetric Channels

Let W_1 and W_2 be two BSCs, each with crossover probability δ. Note that W_1 and W_2 are symmetric [17, p. 185] and have the same capacity-achieving input distribution. We evaluate the bounds presented in Theorem 1, Corollary 4, and Corollary 7 for the CM-DMBC having W_1 and W_2 as its component channels. The bounds are depicted in Fig. 1 for the case $\delta = 0.11$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory



(a) $K = 2$. (b) $K = 3$.

(c) $K = 4$. (d) Normal approximation for $K \in \{2, \cdots, 8\}$.

Fig. 1. Comparison between the achievability bound in Theorem 1, the converse bound in Corollary 4, and the normal approximation (80) for $\epsilon = 10^{-3}$. The component channels in the CM-DMBC are BSCs with crossover probability 0.11. The normal approximation corresponds to the asymptotic expansion in (80) with the $o(\cdot)$ term neglected. The blue curve labeled "achievability, $K = 1$" corresponds to the single-user achievability bound in [8, Th. 3] evaluated for a BSC with crossover probability 0.11 and $\epsilon = 10^{-3}$.

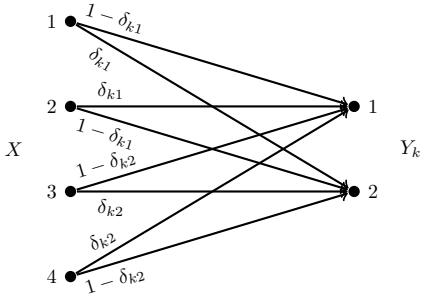


Fig. 2. Asymmetric channels $\{W_k\}$ that obey the conditions in Corollary 9 for the channel parameters $\delta_{11} = 0.01, \delta_{12} = 0.40, \delta_{21} = 0.15$, and $\delta_{22} = 0.10$. The channels consist of two BSCs with common outputs.

$\epsilon = 10^{-3}$. The capacity-achieving input distribution P^* of the individual BSCs is Bern$(1/2)$, their capacity is given by $1 - H_b(\delta)$, where $H_b(\cdot)$ denotes the binary entropy function, and the directional derivatives $\nabla I_k(\cdot)$ of the mutual information at P^* are zero. Furthermore, for $Y_k^n \sim P_{Y_k^n|X^n=x^n}$, the information densities $i_{P^*,W_k}(x^n; Y_k^n)$ satisfy

$$i_{P^*,W_k}(x^n; Y_k^n) \sim n \log(2 - 2\delta) + \log\left(\frac{\delta}{1-\delta}\right) \sum_{j=1}^{n} Z_{k,j} \tag{123}$$

where $Z_{k,j} \overset{\text{i.i.d.}}{\sim} \text{Bern}(\delta)$. We observe that the distribution of the information density in (123) is independent of x^n. The converse bound in the figure is obtained from Corollary 4, where the value of η is optimized numerically. The achievability bound is obtained from Theorem 1 for the choice $P_{X^\infty} = (P^*)^\infty$. To evaluate the bound, we use that τ_1 and τ_2 are i.i.d. RVs. This allows us to compute $\mathbb{E}[\max\{\tau_1, \tau_2\}]$ by evaluating $\sum_{t=0}^{\infty}(1 - F_{\tau_1}(t)^2)$ where $F_{\tau_1}(\cdot)$ is the cumulative distribution function of τ_1. To estimate (20), we use the change of measure technique (see [8, p. 4911]).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

We observe that, in the two-user case, the speed of convergence to the asymptotic limit is indeed slower than for the single-user case (the curve marked "achievability (single)" in Fig. 1, which is the point-to-point achievability bound reported in [8, Th. 3]). In particular, for $\ell \geq 1000$ and $K = 2$, our converse bound is strictly below the single-user achievability bound, which implies that the maximum coding rate for the two-user case is strictly smaller than that for the single-user case. Additionally, the speed of convergence becomes slower as the number of users increases.

B. Asymmetric Channels

Next, we illustrate through an example that Theorem 8 indeed improves over Theorem 5. We consider the CM-DMBC depicted in Fig. 2; we shall also assume that $\delta_{11} = 0.01, \delta_{12} = 0.40, \delta_{21} = 0.15$, and $\delta_{22} = 0.10$. The two component channels, W_1 and W_2, are two BSCs with common outputs. Let now $\mathcal{X}_1 = \{1, 2\}$, $\mathcal{X}_2 = \{3, 4\}$, and $W_{k,r}$ be a BSC with crossover probability $\delta_{k,r}$ for $k \in \{1, 2\}$ and $r \in \{1, 2\}$. One can verify that the condition $P^*(x) + CP^*_{r(x)}(x)\beta'_{r(x)}(w) \in [0, 1]$ for $x \in \mathcal{X}$ and $w \in \mathbb{R}$ in Corollary 9 is satisfied using Lemma 10. Therefore, the asymptotic expansion in (100) holds, i.e., the converse bound in Theorem 5 is tight up to second order. The same is not true for the achievability bound in Theorem 5. Indeed, by computing (53) and (54), we find that $\Xi_{\mathrm{c}} = 0.2630$ but that $\Xi_{\mathrm{a}} = 0.3175$.

V. Conclusion

In this paper, we considered the K-user CM-DMBC for the scenario where variable-length stop-feedback codes are used. We presented achievability and converse bounds on the maximum coding rate $\frac{1}{\ell} \log M^*_{\mathrm{sf}}(\ell, \epsilon)$ for a fixed average blocklength ℓ and average error probability ϵ. The main novelty in our nonasymptotic analysis is the converse bound, which relies on a nonstandard application of the meta-converse theorem [4, Th. 26] to the variable-length setup. The achievability bound follows instead from a straightforward generalization of [8, Th. 3]. An asymptotic analysis of our bounds in the limit $\ell \to \infty$ reveals that, under mild technical conditions, the second-order asymptotic expansion of $\log M^*_{\mathrm{sf}}(\ell, \epsilon)$ contains a square-root penalty. We provided upper and lower bounds on the second-order term in the asymptotic expansion of $\log M^*_{\mathrm{sf}}(\ell, \epsilon)$ and derived necessary and sufficient conditions for our bounds on the second-order term to match. This occurs for example for the case when the component channels are two identical BSCs. For this case, we provide numerical evidence that the convergence to the asymptotic limit of the maximum coding rate is indeed slower than in the point-to-point case. Furthermore, our numerical results show that the first two terms in the asymptotic expansion of $\log M^*_{\mathrm{sf}}(\ell, \epsilon)$ approximate $\log M^*_{\mathrm{sf}}(\ell, \epsilon)$ accordingly.

Finally, we emphasize that our results are based on a setup in which the encoder output at time n does not depend on the stop signals received before time n. A generalization of our analysis to the case when this dependency is allowed (which may result in a faster convergence to capacity) is left for future work. We recently showed that the dispersion is zero if full feedback is available [21]. It is then natural to ask how much feedback is needed for the dispersion to vanish.

Appendix I
Proof of Theorem 1

The proof follows closely [8, Th.3]. Let S be a Bernoulli RV with $\mathbb{P}[S = 1] = q$ and let its probability mass function be given by P_S. We start by specifying $U, f_n, \{g_{k,n}\}_{k\in\mathcal{K}}, \{\tau^*_k\}_{k\in\mathcal{K}}$. The RV U has the following domain and probability mass function

$$\mathcal{U} \triangleq \{0, 1\} \times \underbrace{\mathcal{X}^\infty \times \cdots \times \mathcal{X}^\infty}_{M \text{ times}} \tag{124}$$

$$P_U \triangleq P_S \times \underbrace{P_{X^\infty} \times \cdots \times P_{X^\infty}}_{M \text{ times}}. \tag{125}$$

Note that the cardinality of $\mathcal{U}$ is unbounded. As remarked after Definition 1, the cardinality of $\mathcal{U}$ can always be reduced to $K+1$.

As in [8], the realization u of $U = u$ defines a codebook $\{\mathbf{C}^{(u)}_1, \ldots, \mathbf{C}^{(u)}_M\}$ consisting of M infinite dimensional codewords $\mathbf{C}^{(u)}_j \in \mathcal{X}^\infty$, $j \in \mathcal{M}$. Differently from [8], it also defines the RV S, which we shall use as a source of additional common randomness among the K decoders. The encoder operates as follows:

$$f_n(u, j) \triangleq \mathbf{C}^{(u)}_{j,n}, \qquad u \in \mathcal{U}, j \in \mathcal{M}. \tag{126}$$

Here, $\mathbf{C}^{(u)}_{j,n}$ denotes the nth entry of $\mathbf{C}^{(u)}_j$. To keep the notation compact, we shall omit the superscript (u) in the remaining part of the proof. At time n, decoder k computes the information densities

$$A_{k,n}(j) \triangleq i_{P_{X^n}, W^n_k}(\mathbf{C}^n_j; Y^n_k), \qquad j \in \mathcal{M} \tag{127}$$

where the vector $\mathbf{C}^n_j$ contains the first n entries of $\mathbf{C}_j$. Define the stopping times

$$\tau_k(j) \triangleq \mathbb{1}\{S = 0\} \inf\{n \geq 0 : A_{k,n}(j) \geq \gamma\} \tag{128}$$

and let τ^*_k be the time at which decoder k makes the final decision:

$$\tau^*_k \triangleq \min_{j\in\mathcal{M}} \tau_k(j). \tag{129}$$

The output of decoder k at time τ^*_k is

$$g_{k,\tau^*_k}(U, Y^{\tau^*_k}_k) \triangleq \max\{j \in \mathcal{M} : \tau_k(j) = \tau^*_k\}. \tag{130}$$

When $S = 1$, we have $\tau^*_k = 0$, and hence the decoder $g_{k,\tau^*_k}(U, Y^{\tau^*_k}_k)$ outputs M. The average blocklength is then given by

$$\begin{aligned}
&\mathbb{E}\Big[\max_k \tau^*_k\Big] \\
&\quad = (1 - q)\mathbb{E}\Big[\max_k \tau^*_k | S = 0\Big] && (131)\\
&\quad \leq (1 - q)\frac{1}{M}\sum_{j=1}^{M} \mathbb{E}\Big[\max_k \tau_k(j) | J = j, S = 0\Big] && (132)\\
&\quad = (1 - q)\mathbb{E}[\max \tau_k(1) | J = 1, S = 0] && (133)\\
&\quad = (1 - q)\mathbb{E}\Big[\max_k \tau^{(0)}_k\Big] && (134)
\end{aligned}$$

where the expectation is over J, Y^∞_1, Y^∞_2, and U. Here, (132) follows from (129); (133) follows from symmetry; and (134) fol-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

lows from the definition of $\tau_k^{(0)}$ in (17). For the error probability, we have that

$$\begin{aligned}
&\mathbb{P}\Big[g_{k,\tau_k^*}(U, Y_k^{\tau_k^*}) \neq J\Big] \\
&\quad \leq q + (1-q)\mathbb{P}\Big[g_{k,\tau_k^*}(U, Y_k^{\tau_k^*}) \neq J | S = 0\Big] && (135)\\
&\quad \leq q + (1-q)\mathbb{P}\Big[g_{k,\tau_k^*}(U, Y_k^{\tau_k^*}) \neq 1 | J = 1, S = 0\Big] && (136)\\
&\quad \leq q + (1-q)\mathbb{P}[\tau_k(1) \geq \tau_k^* | S = 0] && (137)\\
&\quad = q + (1-q)\mathbb{P}\left[\bigcup_{j=2}^{M}\{\tau_k(1) \geq \tau_k(j)\}\middle| S = 0\right] && (138)\\
&\quad \leq q + (1-q)(M-1)\mathbb{P}[\tau_k(1) \geq \tau_k(2) | S = 0] && (139)\\
&\quad = q + (1-q)(M-1)\mathbb{P}\Big[\tau_k^{(0)} \geq \bar{\tau}_k^{(0)}\Big]. && (140)
\end{aligned}$$

Here, (136) follows from (130) and (140) follows by noting that, given $J = 1$, the RVs $(A_{k,n}(1), A_{k,n}(2), \tau_k(1), \tau_k(2))$ (where the RVs $\{A_{k,n}(j)\}$ are defined in (127)) have the same joint distribution as $\Big(i_{P_{X^n},W_k^n}(X^n; Y_k^n), i_{P_{X^n},W_k^n}(\bar{X}^n; Y_k^n), \tau_k^{(0)}, \bar{\tau}_k^{(0)}\Big)$. We conclude the proof by noting that, by Definition 1, the tuple $(U, f_n, \{g_{k,n}\}, \{\tau_k^*\})$ defines an (ℓ, M, ϵ)-VLSF code satisfying (19) and (20).

The upper bound in (21) follows from the same steps as in [8, Eqs. (111)–(118)].

Appendix II
Proof of Lemma 2

Let $\epsilon_k^{(u)} \triangleq \mathbb{P}[J \neq g_{k,\tau_k}(U, Y_k^{\tau_k}) | U = u]$, $u \in \mathcal{U}$, be the conditional error probability at decoder k given $U = u$ and define the following probability measure on $\mathcal{X}^\infty \times \overline{\mathscr{Y}}_1^{(u)} \times \cdots \times \overline{\mathscr{Y}}_2^{(u)}$:

$$\mathbb{Q}^{(u)}_{\mathbf{X},\overline{\mathbf{Y}}_1,\cdots,\overline{\mathbf{Y}}_K}(\mathbf{x}, \overline{\mathbf{y}}_1, \cdots, \overline{\mathbf{y}}_K) \triangleq P_{\mathbf{X}}^{(u)}(\mathbf{x}) \prod_{k=1}^{K} Q_k(\overline{\mathbf{y}}_k). \tag{141}$$

For notational convenience, we shall indicate the two probability measures in (28) and (141) simply as $\mathbb{P}^{(u)}$ and $\mathbb{Q}^{(u)}$, respectively. For each decoder k, the average error probability is equal to $\epsilon_k^{(u)}$ under $\mathbb{P}^{(u)}$, and it is no larger than $1 - 1/M$ under $\mathbb{Q}^{(u)}$. Hence, by an application of the meta-converse theorem [4, Th. 26], we conclude that for every $u \in \mathcal{U}$ and $k \in \mathcal{K}$

$$M \leq \frac{1}{\beta_{1-\epsilon_k^{(u)}}(\mathbb{P}^{(u)}_{\mathbf{X},\overline{\mathbf{Y}}_\mathbf{k}}, \mathbb{Q}^{(u)}_{\mathbf{X},\overline{\mathbf{Y}}_\mathbf{k}})}. \tag{142}$$

Here, $\beta_\alpha(P, Q)$ denotes the Neyman-Pearson function which is the minimum type-II error probability of a binary hypothesis test between two probability distributions P and Q on a common measurable space subject to the constraint that the type-I error probability does not exceed $1 - \alpha$. In order to obtain an information spectrum-type bound, we apply the following inequality [4, Eq. (102)]

$$\alpha \leq P\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} \geq \gamma\Big] + \gamma\beta_\alpha(P, Q) \tag{143}$$

to the left-hand side of (142). Here, $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative. By doing so, we find that for every $\gamma_k^{(u)} > 0$ and for every $k \in \mathcal{K}$

$$\log M \leq \log \gamma_k^{(u)} - \log\Big|\mathbb{P}^{(u)}\Big[i_k^{(u)}(\mathbf{X}; \overline{\mathbf{Y}}_k) \leq \log \gamma_k^{(u)}\Big] - \epsilon_k^{(u)}\Big|^+. \tag{144}$$

Here,

$$\begin{aligned}
&i_k^{(u)}(\mathbf{x}; \bar{\mathbf{y}}_k) \\
&\quad \triangleq \log \frac{\mathbb{P}^{(u)}_{\mathbf{X},\overline{\mathbf{Y}}_k}(\mathbf{x}, \overline{\mathbf{y}}_k)}{\mathbb{Q}^{(u)}_{\mathbf{X},\overline{\mathbf{Y}}_k}(\mathbf{x}, \overline{\mathbf{y}}_k)} = \log \frac{\prod_{i=1}^{\mathrm{len}(\overline{\mathbf{y}}_k)} W_k(\overline{y}_{k,i}|x_i)}{Q_k(\overline{\mathbf{y}}_k)} && (145)\\
&\quad = i_k(\mathbf{x}; \overline{\mathbf{y}}_k), \qquad \mathbf{x} \in \mathcal{X}^\infty, \overline{\mathbf{y}}_k \in \overline{\mathscr{Y}}_k^{(u)} && (146)
\end{aligned}$$

where $i_k(\mathbf{x}; \overline{\mathbf{y}}_k)$ was defined in (26). Note that $i_k(\mathbf{x}; \overline{\mathbf{y}}_k)$ depends on $\mathbf{x} \in \mathcal{X}^\infty$ only through its first $\mathrm{len}(\overline{\mathbf{y}}_k)$ entries. Set now

$$\gamma_k^{(u)} \triangleq \sup\Big\{\nu \in \mathbb{R} : \mathbb{P}^{(u)}\big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) \leq \log \nu\big] \leq \epsilon_k^{(u)} + \eta\Big\} \tag{147}$$

where $\eta > 0$ is arbitrary. Using (147), we conclude that

$$\begin{aligned}
&\mathbb{P}^{(u)}\Big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) < \log \gamma_k^{(u)}\Big] \\
&\qquad \leq \epsilon_k^{(u)} + \eta \leq \mathbb{P}^{(u)}\Big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) \leq \log \gamma_k^{(u)}\Big]. && (148)
\end{aligned}$$

Substituting the right-hand side of (148) in (144), we obtain

$$\begin{aligned}
&\log M \\
&\leq \log \gamma_k^{(u)} - \log\Big(\mathbb{P}^{(u)}\Big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) \leq \log \gamma_k^{(u)}\Big] - \epsilon_k^{(u)}\Big) && (149)\\
&\leq \log \gamma_k^{(u)} - \log \eta. && (150)
\end{aligned}$$

Finally, the lemma is establishes by substituting (150) into (148), which yields

$$\begin{aligned}
&\mathbb{P}^{(u)}\big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) < \log(M\eta)\big] \\
&\qquad \leq \mathbb{P}^{(u)}\Big[i_k(\mathbf{X}; \overline{\mathbf{Y}}_k) < \log \gamma_k^{(u)}\Big] && (151)\\
&\qquad \leq \epsilon_k^{(u)} + \eta. && (152)
\end{aligned}$$

Appendix III
Proof of Theorem 3 (cardinality bound)

We simplify the minimization problem in (45) by showing that the minimum is also attained under the additional constraint that $|\overline{\mathcal{U}}| \leq K + 1$. Define the $(K+1)$-dimensional region

$$\mathcal{R} \triangleq \Big\{(\varepsilon_1, \cdots, \varepsilon_K, L) \in [0,1]^K \times \mathbb{R}_+ : L \geq \sum_{t=0}^{\infty}\left(1 - L_t\left(\varepsilon_1, \cdots, \varepsilon_K\right)\right)\Big\}. \tag{153}$$

Furthermore, let $\mathcal{R}_{\text{convex}}$ be the convex hull of $\mathcal{R}$. Suppose that $\mathbf{p}_0$ lies on the lower convex envelope of $\mathcal{R}_{\text{convex}}$. Then we can write $\mathbf{p}_0$ as a convex combination of $I \in \mathbb{N}$ points in $\mathcal{R}$:

$$\mathbf{p}_0 = \sum_{i=1}^{I} \alpha_i \mathbf{p}_i \tag{154}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

where $\mathbf{p}_i \in \mathcal{R}$, $\alpha_i > 0$, and $\sum_{i=1}^{I} \alpha_i = 1$. Since $\mathbf{p}_0$ is a boundary point of $\mathcal{R}_{\text{convex}}$, there exists a supporting hyperplane $\{\mathbf{p} \in \mathbb{R}^{K+1} : \mathbf{a}^T\mathbf{p} = \mathbf{a}^T\mathbf{p}_0\}$ for some $\mathbf{a} \neq \mathbf{0}_{K+1}$, $\mathbf{a} \in \mathbb{R}^{K+1}$, with the property that $\mathbf{a}^T\mathbf{p} \leq \mathbf{a}^T\mathbf{p}_0$ for every $\mathbf{p} \in \mathcal{R}_{\text{convex}}$ [22, pp. 50–51]. Now we note that the points $\{\mathbf{p}_i\}$, $i \in \{1, \cdots, I\}$, must be on the supporting hyperplane $\{\mathbf{p} \in \mathbb{R}^{K+1} : \mathbf{a}^T\mathbf{p} = \mathbf{a}^T\mathbf{p}_0\}$ for every $i \in \{1, \cdots, I\}$. Indeed, suppose on the contrary that $\mathbf{a}^T\mathbf{p}_i < \mathbf{a}^T\mathbf{p}_0$ for some $i \in \{1, \cdots, I\}$. Then we have a contradiction:

$$\mathbf{a}^T\mathbf{p}_0 = \sum_{i=1}^{I} \alpha_i \mathbf{a}^T\mathbf{p}_i < \sum_{i=1}^{I} \alpha_i \mathbf{a}^T\mathbf{p}_0 = \mathbf{a}^T\mathbf{p}_0. \quad (155)$$

Now, define the region

$$\mathcal{R}_0 \triangleq \mathcal{R} \cap \Big\{\mathbf{p} = (p_1, \ldots, p_{K+1}) \in \mathbb{R}^{K+1} : \mathbf{a}^T\mathbf{p} = \mathbf{a}^T\mathbf{p}_0 \text{ and } p_{K+1} \leq \max_{1 \leq i \leq I} p_{i,K+1}\Big\} \quad (156)$$

and the convex hull $\mathcal{R}_{0,\text{convex}}$ of $\mathcal{R}_0$. Here, $p_{i,K+1}$ denotes the $(K+1)$th entry of $\mathbf{p}_i$. The region $\mathcal{R}_0$ is closed because $L_t(\cdot)$ is continuous in $\{\varepsilon_k^{(u)}\}$ and, since the $(K+1)$th entry of every point in $\mathcal{R}_0$ is bounded from above by $\max_{1 \leq i \leq I} p_{i,K+1}$, the region is bounded in $\mathbb{R}^{K+1}$. This implies that $\mathcal{R}_0$ is compact. Moreover, $\mathbf{p}_0 \in \mathcal{R}_{0,\text{convex}}$ because $\mathbf{p}_1, \cdots, \mathbf{p}_I \in \mathcal{R}_0$, and $\mathcal{R}_0$ lies in a K-dimensional affine subspace of $\mathbb{R}^{K+1}$. Hence, Caratheodory theorem [17, Th. 15.3.5] implies that $\mathbf{p}_0$ can be written as a convex combination of at most $K+1$ points in $\mathcal{R}_0$. But since $\mathcal{R}_0 \subseteq \mathcal{R}$, we can also write $\mathbf{p}_0$ as a convex combination of at most $K+1$ points in $\mathcal{R}$.

Now, observe that the point

$$\left(\mathbb{E}\left[\varepsilon_1^{(\bar{U})}\right], \cdots, \mathbb{E}\left[\varepsilon_K^{(\bar{U})}\right], \mathbb{E}\left[\sum_{t=0}^{\infty}\left(1 - L_t\left(\varepsilon_1^{(\bar{U})}, \cdots, \varepsilon_K^{(\bar{U})}\right)\right)\right]\right) \quad (157)$$

evaluated for the RV $\bar{U}$ supported on the set $\bar{\mathcal{U}}$ and distributed as $P_{\bar{U}}$, and for the $\{\varepsilon_k^{(\bar{u})}\}$ that minimize (45), is a boundary point of $\mathcal{R}_{\text{convex}}$. By the above argument, we conclude that (45) is equal to

$$\min_{\substack{P_{\bar{U}} \in \mathcal{P}(\bar{\mathcal{U}}), \varepsilon_k^{\bar{u}} \in [0,1]: \\ \mathbb{E}\left[\varepsilon_k^{(\bar{U})}\right] \leq \epsilon + \eta}} \mathbb{E}\left[\sum_{t=0}^{\infty}\left(1 - L_t\left(\varepsilon_1^{(\bar{U})}, \cdots, \varepsilon_K^{(\bar{U})}\right)\right)\right] \quad (158)$$

where the RV $\bar{U}$ is supported on the set $\bar{\mathcal{U}}$ with $|\bar{\mathcal{U}}| \leq K+1$.

Appendix IV
Proof of Theorem 5 (converse)

Fix a family of (ℓ, M, ϵ)-VLSF codes parameterized by the blocklength ℓ. We shall assume that $\liminf_{\ell \to \infty} \log(M)/\ell > 0$, that is, M grows at least exponentially with ℓ. If this does not occur, then the rightmost inequality in (52) holds trivially. To establish the desired result, we analyze the nonasymptotic converse bound in Theorem 3 in the limit $\log M \to \infty$. We shall set $\eta = (\log M)^{-1}$ and choose the auxiliary distributions $\{Q_k^{(\infty)}\}$ as follows. Let x^t, $t \in \mathbb{N}$, be an arbitrary t-dimensional vector in $\mathcal{X}^t$ and let $P_{x^t} \in \mathcal{P}(\mathcal{X})$ denote its type [23, Def. 2.1]. Furthermore, let $Q_{k,x^t}^{(\infty)}$ be the product distribution on $\mathcal{Y}_k^{\infty}$ generated by the marginal distribution $P_{x^t}W_k$. Finally, let $\mathcal{P}_t(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$ be the set of types of t-dimensional sequences. We choose $Q_k^{(\infty)}$ as follows:

$$Q_k^{(\infty)}(\mathbf{y}) = \sum_{t=1}^{\lfloor \frac{2}{C}\log M \rfloor} \sum_{\substack{x^t \in \mathcal{X}^t: \\ P_{x^t} \in \mathcal{P}_t(\mathcal{X})}} \frac{Q_{k,x^t}^{(\infty)}(\mathbf{y})}{\lfloor \frac{2}{C}\log M \rfloor |\mathcal{P}_t(\mathcal{X})|}. \quad (159)$$

Here, the inner sum is taken over the set of types of t-dimensional sequences and C is the channel capacity given in (1). To keep notation compact, we set

$$\tilde{\imath}_k(x^n; y^n) \triangleq i_{P_{x^t}, W_k}(x^n; y^n) \quad (160)$$

which is defined for $n \leq t$ and for every $x^t \in \mathcal{X}^t$ and $y^n \in \mathcal{Y}_k^n$.

We shall next summarize the key steps of the proof. These steps are analyzed in details in Sections IV-A–IV-D.

Step 1: We obtain an upper bound on $L_t(\boldsymbol{\varepsilon})$ in Theorem 3 that does not involve any maximization over n (the inner maximization in (29)). Specifically, we show in Appendix IV-A that, whenever $t \leq \lfloor \frac{2}{C}\log M \rfloor$, we can dispose of this inner maximization by adding an error term of order $1/\log M$:

$$L_t(\boldsymbol{\varepsilon}) \leq \max_{x^t \in \mathcal{X}^t} \prod_k \min\left\{1, \mathbb{P}\left[\tilde{\imath}_k(x^t; Y_k^t) \geq \lambda\right] + \varepsilon_k\right\} + \frac{2^K - 1}{\lambda}. \quad (161)$$

Here,

$$\lambda \triangleq \log M - 2\log\log M - |\mathcal{X}| \log\left(\frac{2}{C}\log M + 1\right). \quad (162)$$

and we denoted $L_t(\varepsilon_1, \cdots, \varepsilon_K)$ by $L_t(\boldsymbol{\varepsilon})$ with $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_K]$.

Step 2: We use (161) to lower-bound the right-hand side of (30) in Theorem 3. Since (161) holds only for $t \leq \lfloor \frac{2}{C}\log M \rfloor$, we must truncate the infinite sum in (30) as follows:

$$\sum_{t=0}^{\infty}(1 - L_t(\boldsymbol{\varepsilon})) \geq \sum_{t=0}^{\beta_{\boldsymbol{\varepsilon}}}(1 - L_t(\boldsymbol{\varepsilon})) \quad (163)$$

where $\beta_{\boldsymbol{\varepsilon}} \in \mathbb{N}$ is given by

$$\beta_{\boldsymbol{\varepsilon}} \triangleq \left\lfloor \frac{\lambda}{C} + \sqrt{\frac{\lambda V}{C^3}}\nu_{\boldsymbol{\varepsilon}} \right\rfloor. \quad (164)$$

Here, V is defined in Theorem 5 and $\nu_{\boldsymbol{\varepsilon}}$ is the solution of

$$\prod_k [Q(-\varrho_k \nu_{\boldsymbol{\varepsilon}}) + (1 - 2\delta_1)\varepsilon_k + \delta_1] = 1 \quad (165)$$

with $\delta_1 \in (0, 1/2)$ being an arbitrary constant that does not depend on λ. The role of δ_1 is to ensure that $\nu_{\boldsymbol{\varepsilon}}$ is bounded from above and from below for all $\boldsymbol{\varepsilon} \in [0,1]^K$. We note that $\beta_{\boldsymbol{\varepsilon}} \leq \frac{2}{C}\log M$ for sufficiently large M and for every $\boldsymbol{\varepsilon} \in [0,1]^K$. Note that we define $\beta_{\boldsymbol{\varepsilon}}$ as in (164) instead of setting it equal $2/C \log M$ in order to control the error term originating from central limit theorem as we shall see later (see (287)). Hence, since $\beta_{\boldsymbol{\varepsilon}} \leq \frac{2}{C}\log M$, we can use (161) to further lower-bound (163). Note that $M \to \infty$ implies $\lambda \to \infty$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Step 3: Next, we characterize the asymptotic behavior of the upper bound (161) in the limit $M \to \infty$. This will be used to provide an asymptotic lower bound on the right-hand side of (163). It turns out convenient to subdivide the interval $[0, \beta_{\boldsymbol{\varepsilon}}]$ into $K+2$ subintervals and to perform a different asymptotic analysis on each of the subintervals. Specifically, we set $[0, \beta_{\boldsymbol{\varepsilon}}] = \bigcup_{i=1}^{K+1} \mathcal{T}_i$ where $\mathcal{T}_i = [t_i, t_{i+1})$, $i \in \{0, \cdots, K\}$, and $\mathcal{T}_{K+1} = [t_{K+1}, \beta_{\boldsymbol{\varepsilon}}]$ with $t_0 = 0$ and

$$t_i \triangleq \left\lfloor \frac{\lambda}{C_i} - \sqrt{\frac{V\lambda}{C^3}} \log \lambda \right\rfloor, \quad i \in \{1, \cdots, K\} \tag{166}$$

$$t_{K+1} \triangleq \left\lfloor \frac{\lambda}{C} - \sqrt{\frac{V\lambda}{C^3}} \log \lambda \right\rfloor. \tag{167}$$

Recall that since $C_1 \geq C_2 \geq \cdots \geq C_K \geq C$ by assumption, we have that $t_0 \leq t_1 \leq \cdots \leq t_{K+1}$. Additionally, for sufficiently large M, we also have that $t_{K+1} < \beta_{\boldsymbol{\varepsilon}}$.

In the first $K+1$ subintervals, we upper-bound (161) by means of a large-deviation analysis based on Hoeffding's inequality (see Appendix IV-B). In the last interval, our upper bound relies on Chebyshev's inequality and the Berry-Esseen central limit theorem (see Appendix IV-C). These bounds are used to further lower-bound the right-hand side of (163), as illustrated in Fig. 3.

We next summarize the asymptotic behavior of the bounds obtained in Sections IV-B–IV-C.

To do so, it is convenient to introduce some notation that will allow us to keep our expressions compact. First, let $\rho > 0$ and $\delta_2 \in (0, 1)$ be constants. For reasons that will become apparent later, we need ρ to satisfy

$$\frac{1}{C_i} + \frac{i}{C_K} - \frac{(i+1)\rho}{C_1} > \frac{i}{C}, \quad i \in \{1, \cdots, K-1\}. \tag{168}$$

Note that the assumption (51) ensures that one can find a ρ that satisfies (168). Let

$$d_0 \triangleq \frac{1}{C} - \frac{1}{C_K} + \frac{\rho}{C_1} \tag{169}$$

$$d_1 \triangleq \frac{1}{C_1}(1-\rho) \tag{170}$$

$$d_i \triangleq \frac{1}{C_i} - \frac{1}{C_{i-1}}, \qquad i \in \{2, \cdots, K\} \tag{171}$$

and define, for $\boldsymbol{\varepsilon} \in [0,1]^K$, the function[8]

$$f(\boldsymbol{\varepsilon}) \triangleq \frac{1}{C} - d_0 \left(\max_k \varepsilon_k \right) - \sum_{i=1}^{K} d_i \left(\prod_{k \in \{i, \cdots, K\}} \varepsilon_k \right). \tag{172}$$

This function is continuous in $\boldsymbol{\varepsilon} \in [0,1]^K$ and satisfies $f(\mathbf{0}_K) = 1/C$ and $f(\mathbf{1}_K) = 0$. We shall also need the function $g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon})$ given in (291) (its exact expression is not important for the level of detail provided in this section). This function is continuous in $\boldsymbol{\varepsilon} \in [0,1]^K$, $\delta_1 > 0$, and $\delta_2 > 0$, and has the following limits:

$$\lim_{\delta_1 \to 0, \delta_2 \to 0} g_{\delta_1,\delta_2}(\mathbf{0}_K) = \sqrt{\frac{V}{C^3}} \mathbb{E}\left[\max_k H_k \right] \tag{173}$$

$$\lim_{\delta_1 \to 0, \delta_2 \to 0} g_{\delta_1,\delta_2}(\mathbf{1}_K) = 0. \tag{174}$$

[8]We use the convention that $\prod_{k \in \emptyset} a_k = 1$ for an arbitrary sequence $\{a_k\}$.

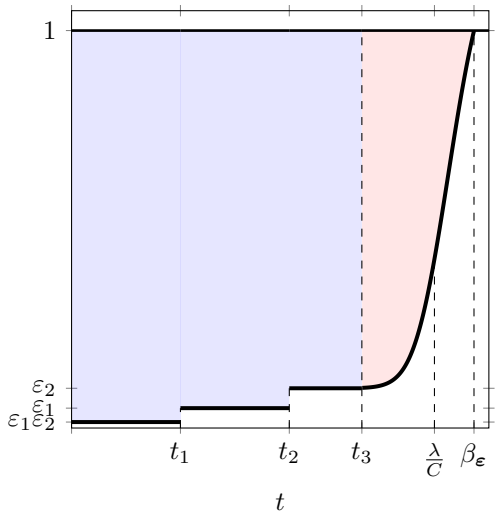


Fig. 3. Our approach to lower-bounding (163). For the case $K = 2$ and $\boldsymbol{\varepsilon} = (0.05, 0.10)$, the range of t is divided into four subintervals: $[0, t_1)$, $[t_1, t_2)$, $[t_2, t_3)$, and $[t_3, \beta_{\boldsymbol{\varepsilon}}]$. The area of the blue shaded region depicts $\sum_{t=0}^{t_3-1}(1 - L_t(\boldsymbol{\varepsilon}))$ while the area of the red shaded region depicts $\sum_{t=t_3}^{\beta_{\boldsymbol{\varepsilon}}}(1 - L_t(\boldsymbol{\varepsilon}))$. The plotted curve represents a nonasymptotic upper bound on $L_t(\boldsymbol{\varepsilon})$ that is provided in Appendix IV-B and Appendix IV-C.

In Appendix IV-B, we prove the following asymptotic bound, which holds for every $\boldsymbol{\varepsilon} \in [0,1]^K$ and for sufficiently large λ:

$$\sum_{t=0}^{t_{K+1}-1} (1 - L_t(\boldsymbol{\varepsilon})) \geq \lambda f(\boldsymbol{\varepsilon}) - \sqrt{\frac{\lambda V}{C^3}} \log(\lambda) \left(1 - \max_k \varepsilon_k \right) - \mathbb{c}. \tag{175}$$

This bound holds for every $\boldsymbol{\varepsilon} \in [0,1]^K$. Furthermore, in Appendix IV-C, we provide the following asymptotic bound:

$$\sum_{t=t_{K+1}}^{\beta_{\boldsymbol{\varepsilon}}} (1 - L_t(\boldsymbol{\varepsilon})) \geq \sqrt{\lambda} g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}) + \sqrt{\frac{\lambda V}{C^3}} \log(\lambda) \left(1 - \max_k \varepsilon_k \right) + \mathcal{O}(\log \lambda). \tag{176}$$

Here, the $\mathcal{O}(\log \lambda)$ term is uniform in $\boldsymbol{\varepsilon} \in [0,1]^K$.

By combining (163), (175), and (176), we obtain for all $\boldsymbol{\varepsilon} \in [0,1]^K$ and for all sufficiently large λ

$$\sum_{t=0}^{\infty} (1 - L_t(\boldsymbol{\varepsilon})) \geq \sum_{t=0}^{\beta_{\boldsymbol{\varepsilon}}} (1 - L_t(\boldsymbol{\varepsilon})) \tag{177}$$

$$\geq \lambda f(\boldsymbol{\varepsilon}) - \sqrt{\frac{\lambda V}{C^3}} \log(\lambda)(1 - \max_k \varepsilon_k) - \mathbb{c} + \sqrt{\lambda} g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}) + \sqrt{\frac{\lambda V}{C^3}} \log(\lambda) \left(1 - \max_k \varepsilon_k \right) + \mathcal{O}(\log \lambda) \tag{178}$$

$$= \lambda f(\boldsymbol{\varepsilon}) + \sqrt{\lambda} g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}) + \mathcal{O}(\log \lambda). \tag{179}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Again, we note that the $\mathcal{O}(\log\lambda)$ term in (179) is uniform in $\boldsymbol{\varepsilon}$.

Step 4: We are left with solving the minimization in (30). Specifically, we need to evaluate $\min\left\{\lambda f(\boldsymbol{\varepsilon}^{(U)}) + \sqrt{\lambda} g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}^{(U)})\right\}$, where the minimization is over all $\{\boldsymbol{\varepsilon}^{(u)}\}_{u\in\mathcal{U}}$ and all probability distributions $P_U \in \mathcal{P}(\mathcal{U})$ subject to $\mathbb{E}_U\left[\boldsymbol{\varepsilon}^{(u)}\right] \le \epsilon + (\log M)^{-1}$. To do so, we rely on [4, Lem. 63] which is repeated here for convenience.

Lemma 11 ([4, Lem. 63]): Let D be a compact metric space. Suppose $f : D \mapsto \mathbb{R}$ and $g : D \mapsto \mathbb{R}$ are continuous. Define

$$f^* \triangleq \max_{x\in D} f(x) \tag{180}$$

and

$$g^* \triangleq \sup_{x: f(x)=f^*} g(x). \tag{181}$$

Then,

$$\max_{x\in D}\left[nf(x) + \sqrt{n}g(x)\right] = nf^* + \sqrt{n}g^* + o(\sqrt{n}). \tag{182}$$

As a first step, we show in Appendix IV-D that

$$\min_{\substack{P_U\in\mathcal{P}(\mathcal{U}),\boldsymbol{\varepsilon}^{(u)}\in[0,1]^K:\\ \mathbb{E}_U\left[\varepsilon_k^{(U)}\right]\le\epsilon+(\log M)^{-1}}} \mathbb{E}_U\left[f(\boldsymbol{\varepsilon}^{(U)})\right] = \frac{1-\epsilon-(\log M)^{-1}}{C}. \tag{183}$$

and that the set of minimizers of the left-hand side of (183) is given by

$$\mathcal{G} \triangleq \left\{\left(P_U, \{\boldsymbol{\varepsilon}^{(u)}\}\right) : \mathbb{E}_U\left[\varepsilon_k^{(U)}\right] = \epsilon + (\log M)^{-1} \text{ and } P_U(u) > 0 \Rightarrow \boldsymbol{\varepsilon}^{(u)} \in \{\mathbf{0}_K\}\cup\{\mathbf{1}_K\} \text{ for } u\in\mathcal{U}\right\}. \tag{184}$$

Next, it follows from (184) that

$$\begin{aligned}
&\min_{(P_U,\{\boldsymbol{\varepsilon}^{(u)}\})\in\mathcal{G}} \mathbb{E}_U\left[g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}^U)\right] \\
&= \min_{(P_U,\{\boldsymbol{\varepsilon}^{(u)}\})\in\mathcal{G}} \sum_{u\in\mathcal{U}: P_U(u)>0} P_U(u) g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}^u) && (185)\\
&= \min_{(P_U,\{\boldsymbol{\varepsilon}^{(u)}\})\in\mathcal{G}} \Bigg[g_{\delta_1,\delta_2}(\mathbf{0}_K) \sum_{\substack{u\in\mathcal{U}:\\ P_U(u)>0,\boldsymbol{\varepsilon}^{(u)}=\mathbf{0}_K}} P_U(u) \\
&\qquad + g_{\delta_1,\delta_2}(\mathbf{1}_K) \sum_{\substack{u\in\mathcal{U}:\\ P_U(u)>0,\boldsymbol{\varepsilon}^{(u)}=\mathbf{1}_K}} P_U(u)\Bigg] && (186)\\
&= g_{\delta_1,\delta_2}(\mathbf{0}_K)\left(1-\epsilon-(\log M)^{-1}\right) \\
&\quad + g_{\delta_1,\delta_2}(\mathbf{1}_K)(\epsilon+(\log M)^{-1}). && (187)
\end{aligned}$$

By (30) and (179), we have that

$$\begin{aligned}
\ell \ge \min_{\substack{P_U,\boldsymbol{\varepsilon}^{(u)}:\\ \mathbb{E}_U\left[\varepsilon_k^{(U)}\right]\le\epsilon+(\log M)^{-1}}} &\mathbb{E}_U\left[\lambda f(\boldsymbol{\varepsilon}^{(U)}) + \sqrt{\lambda} g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}^{(U)})\right] \\
&+ \mathcal{O}(\log\lambda) && (188)
\end{aligned}$$

where the $\mathcal{O}(\log\lambda)$ term is uniform in $\boldsymbol{\varepsilon}^{(u)}$. We observe that the set of minimizers in (184) and the Euclidean norm form a compact metric space. It follows from Lemma 11 that

$$\begin{aligned}
\ell \ge \min_{\substack{P_U,\boldsymbol{\varepsilon}_1^{(u)},\boldsymbol{\varepsilon}_2^{(u)}:\\ \mathbb{E}_U\left[\varepsilon_k^{(U)}\right]\le\epsilon+(\log M)^{-1}}} &\mathbb{E}_U\left[\lambda f(\boldsymbol{\varepsilon}^{(U)}) + \sqrt{\lambda} g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}^{(U)})\right] \\
&+ \mathcal{O}(\log\lambda) && (189)\\
= \frac{\lambda(1-\epsilon-(\log M)^{-1})}{C} &\\
&+ \sqrt{\lambda}\min_{(P_U,\{\boldsymbol{\varepsilon}\})\in\mathcal{G}} \mathbb{E}_U\left[g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}^{(U)})\right] + \mathcal{O}(\log\lambda) && (190)\\
= \frac{\lambda(1-\epsilon-(\log M)^{-1})}{C} &+ \sqrt{\lambda} g_{\delta_1,\delta_2}(\mathbf{0}_K)\left(1-\epsilon-\frac{1}{\log M}\right) \\
&+ \sqrt{\lambda} g_{\delta_1,\delta_2}(\mathbf{1}_K)\left(\epsilon+(\log M)^{-1}\right) + \mathcal{O}(\log\lambda) && (191)\\
= \frac{\lambda(1-\epsilon)}{C} &+ \sqrt{\lambda} g_{\delta_1,\delta_2}(\mathbf{0}_K)(1-\epsilon) + \sqrt{\lambda} g_{\delta_1,\delta_2}(\mathbf{1}_K)\epsilon \\
&+ \mathcal{O}(\log\lambda). && (192)
\end{aligned}$$

Here, (191) follows from (183) and in (192) we have used that $\lambda(\log M)^{-1} = \mathcal{O}(1)$. Recall that $g_{\delta_1,\delta_2}(\mathbf{0}_K)$ and $g_{\delta_1,\delta_2}(\mathbf{1}_K)$ are continuous in $\delta_1 > 0$ and $\delta_2 > 0$, and that we have the limits (173) and (174). Hence, by choosing δ_1 and δ_2 arbitrarily small, we obtain the inequality

$$\ell \ge \frac{\lambda(1-\epsilon)}{C} + (1-\epsilon)\sqrt{\frac{\lambda V}{C^3}}\mathbb{E}\left[\max_k H_k\right] + o(\sqrt{\lambda}). \tag{193}$$

Here, recall that $\{H_k\}$ have cumulative distribution function given in (55). Finally, using the definition of λ in (162), we conclude that

$$\log M \le \frac{\ell C}{1-\epsilon} - \sqrt{\frac{\ell V}{1-\epsilon}}\mathbb{E}\left[\max_k H_k\right] + o(\sqrt{\ell}) \tag{194}$$

which establishes the desired result.

A. Disposing of the maximum in (29)

By (22), we have that for all $\bar{\mathbf{y}} \in \mathscr{Y}_k$,

$$\begin{aligned}
&Q_k(\bar{\mathbf{y}}) \\
&= \sum_{\substack{\mathbf{y}\in\mathcal{Y}_k^\infty:\\ \bar{\mathbf{y}}=[y_1,\cdots,y_{\mathsf{len}(\bar{y})}]}} \sum_{t=1}^{\lfloor\frac{2}{C}\log M\rfloor} \sum_{P_{x^t}\in\mathcal{P}_t(\mathcal{X})} \frac{Q_{k,x^t}^{(\infty)}(\mathbf{y})}{\lfloor\frac{2}{C}\log M\rfloor|\mathcal{P}_t(\mathcal{X})|} && (195)\\
&= \sum_{t=1}^{\lfloor\frac{2}{C}\log M\rfloor} \sum_{P_{x^t}\in\mathcal{P}_t(\mathcal{X})} \frac{1}{\lfloor\frac{2}{C}\log M\rfloor|\mathcal{P}_t(\mathcal{X})|} \prod_{i=1}^{\mathsf{len}(\bar{\mathbf{y}})} P_{x^t}W_k(\bar{y}_i). && (196)
\end{aligned}$$

Let now $\tilde{t}$ be an integer no larger than $\frac{2}{C}\log M$. Using that Q_k is a convex combination of measures on $\mathscr{Y}_k$, we obtain the following relation between $i_k(x^t; y_k^t)$ and $i_{P_{x^{\tilde{t}}},W_k}(x^t; y_k^t)$

$$\begin{aligned}
&i_k(x^t; y_k^t) \\
&\quad = \log\frac{W_k^t(y_k^t|x^t)}{Q_k(y_k^t)} && (197)\\
&\quad \le \log\frac{W_k^t(y_k^t|x^t)}{\frac{1}{\lfloor\frac{2}{C}\log M\rfloor|\mathcal{P}_{\tilde{t}}(\mathcal{X})|}P_{x^{\tilde{t}}}W_k^t(y_k^t)} && (198)\\
&\quad \le i_{P_{x^{\tilde{t}}},W_k}(x^t; y_k^t) + |\mathcal{X}|\log\left(\frac{2}{C}\log M + 1\right). && (199)
\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

In (198), $P_{x^{\bar{t}}}W_k^t(y_k^t) \triangleq (P_{x^{\bar{t}}}W_k)^t(y_k^t)$ denotes the product distributions induced on $\mathcal{Y}_k^t$ by the probability distribution $(P_{x^{\bar{t}}})^t$. The inequality in (198) follows because the logarithm is monotonically increasing and because the $\{P_{x^{\bar{t}}}W_k^t(y_k^t)\}$ are nonnegative. Finally, (199) follows because $|\mathcal{P}_t(\mathcal{X})| \leq (t+1)^{|\mathcal{X}|}$ [17, Th. 11.1.1]. We can now upper-bound $L_t(\boldsymbol{\varepsilon})$ in (29) for $t \leq \frac{2}{C}\log M$, where $\boldsymbol{\varepsilon} \triangleq [\varepsilon_1, \cdots, \varepsilon_K]$, as follows. Let $\tilde{\lambda}$ be defined as

$$\tilde{\lambda} \triangleq \log M - \log\log M - |\mathcal{X}| \log\left(\frac{2}{C}\log M + 1\right). \quad (200)$$

Then,

$$\begin{aligned}
&L_t(\boldsymbol{\varepsilon}) \\
&= \max_{x^t \in \mathcal{X}^t} \prod_k \min\Bigg\{1, \\
&\qquad \mathbb{P}\left[\max_{0\leq n\leq t} i_k(x^n; Y_k^n) \geq \log M + \log\eta\right] + \varepsilon_k\Bigg\} \quad (201) \\
&\leq \max_{x^t \in \mathcal{X}^t} \prod_k \min\Bigg\{1, \\
&\qquad \mathbb{P}\left[\max_{0\leq n\leq t} i_{P_{x^t},W_k}(x^n; Y_k^n) \geq \tilde{\lambda}\right] + \varepsilon_k\Bigg\} \quad (202) \\
&= \max_{x^t \in \mathcal{X}^t} \prod_k \min\left\{1, \mathbb{P}\left[\max_{0\leq n\leq t} \tilde{\imath}_k(x^n; Y_k^n) \geq \tilde{\lambda}\right] + \varepsilon_k\right\}. \quad (203)
\end{aligned}$$

In (202), we have used (199) and that $\eta = (\log M)^{-1}$, and in (203), we have used (160).

Fix a positive constant ν. We dispose of the inner maximization in (203) through the steps (204)–(207), shown in the top of the next page. In (205), we denoted the last $t-n$ entries of Y_k^t by $Y_{k,n+1}^t$. The inequality in (205) holds because $\tilde{\imath}_k(x^t; Y_k^t) \geq \tilde{\lambda} - \nu$ and $\tilde{\imath}_k(x_{n+1}^t; Y_{k,n+1}^t) \leq -\nu$ imply that $\tilde{\imath}_k(x^n; Y_k^n) = \tilde{\imath}_k(x^t; Y_k^t) - \tilde{\imath}_k(x_{n+1}^t; Y_{k,n+1}^t) \geq \tilde{\lambda}$ for $n \in \{0, \cdots, t-1\}$. Finally, we have used the union bound in (207). Define now

$$\begin{aligned}
&\tilde{\tau}_k(x^t, y^t) \\
&\quad \triangleq \max\big\{\{0\} \cup \{1 \leq n \leq t : \tilde{\imath}_k(x_n^t; y_n^t) \leq -\nu\}\big\}. \quad (208)
\end{aligned}$$

This definition implies the following: if $\tilde{\tau}(x^t, y^t) = n$, then for all sequences $\tilde{y}^t \in \mathcal{Y}_k^t$ such that $\tilde{y}_n^t = y_n^t$, we also have that $\tilde{\tau}_k(x^t, \tilde{y}^t) = n$.

Let $\bar{y}_k^t \in \mathcal{Y}_k^t$ be arbitrary vectors for $k \in \mathcal{K}$. We can now upper-bound the second term in (207) as follows

$$\begin{aligned}
&\mathbb{P}\left[\bigcup_{n=1}^t \{\tilde{\imath}_k(x_n^t; Y_{k,n}^t) \leq -\nu\}\right] \\
&= \sum_{n=1}^t \sum_{\substack{y^t \in \mathcal{Y}_k^t: \\ \tilde{\tau}_k(x^t,y^t)=n}} W_k^t(y^t|x^t) \quad (209) \\
&= \sum_{n=1}^t \sum_{\substack{y^t \in \mathcal{Y}_k^t: \\ \tilde{\tau}_k(x^t,y^t)=n}} \frac{\prod_{i=n}^t W_k(y_i|x_i)}{\prod_{i=n}^t P_{x^t}W_k(y_i)} \frac{\prod_{i=n}^t P_{x^t}W_k(y_i)}{\prod_{i=n}^t W_k(y_i|x_i)} W_k^t(y^t|x^t) \quad (210) \\
&\leq \exp(-\nu) \sum_{n=1}^t \sum_{\substack{y^t \in \mathcal{Y}_k^t: \\ \tilde{\tau}_k(x^t,y^t)=n}} \left(\prod_{i=n}^t P_{x^t}W_k(y_i)\right) \\
&\qquad \times W_k^{n-1}(y^{n-1}|x^{n-1}) \quad (211) \\
&= \exp(-\nu) \sum_{n=1}^t \sum_{\substack{y^t \in \mathcal{Y}_k^t: \\ \bar{y}_k^{n-1}=y^{n-1} \\ \tilde{\tau}_k(x^t,y^t)=n}} \prod_{i=n}^t P_{x^t}W_k(y_i) \quad (212) \\
&= \exp(-\nu) \sum_{n=1}^t P_{x^t}W_k^t[\tilde{\tau}_k(x^t, Y^t) = n] \quad (213) \\
&\leq \exp(-\nu). \quad (214)
\end{aligned}$$

Here, (211) holds because of (160) and because $\tilde{\imath}_k(x_n^t; y_n^t) \leq -\nu$ for every y^t such that $\tilde{\tau}_k(x^t, y^t) = n$; (212) holds because $\{\bar{y}_k^t\}_{k\in\mathcal{K}}$ are fixed arbitrary vectors and because all sequences $\tilde{y}^t$ whose last $t-n+1$ entries coincide with the ones of y^t also satisfy $\tilde{\tau}_k(x^t, \tilde{y}^t) = n$.

By substituting (214) into (207), by choosing $\nu = \log\log M$, and by recalling the definition of λ in (162), we conclude that

$$\begin{aligned}
&\mathbb{P}\left[\max_{0\leq n\leq t} \tilde{\imath}_k(x^n; Y_k^n) \geq \tilde{\lambda}\right] \\
&\qquad \leq \mathbb{P}\big[\tilde{\imath}_k(x^t; Y_k^t) \geq \lambda\big] + \frac{1}{\log M}. \quad (215)
\end{aligned}$$

It follows from (203), from (215), and from the inequality $(\log M)^{-1} \leq 1/\lambda$ that for all $t \leq \frac{2}{C}\log M$, we have

$$\begin{aligned}
&L_t(\boldsymbol{\varepsilon}) \\
&\leq \max_{x^t \in \mathcal{X}^t} \prod_k \min\left\{1, \mathbb{P}\big[\tilde{\imath}_k(x^t; Y_k^t) \geq \lambda\big] + \frac{1}{\log M} + \varepsilon_k\right\} \quad (216) \\
&\leq \max_{x^t \in \mathcal{X}^t} \prod_k \left(\min\big\{1, \mathbb{P}\big[\tilde{\imath}_k(x^t; Y_k^t) \geq \lambda\big] + \varepsilon_k\big\} + \frac{1}{\lambda}\right) \quad (217) \\
&\leq \max_{x^t \in \mathcal{X}^t} \prod_k \min\big\{1, \mathbb{P}\big[\tilde{\imath}_k(x^t; Y_k^t) \geq \lambda\big] + \varepsilon_k\big\} + \frac{2^K - 1}{\lambda}. \quad (218)
\end{aligned}$$

In (218), we have expanded the product in (217) into 2^K terms and used that $\min\{1, \mathbb{P}[\tilde{\imath}_k(x^t; Y_k^t) \geq \lambda]\} \leq 1$. This establishes (161).

B. Large-deviation analysis

To apply Hoeffding's inequality for all t within the first $K+1$ intervals, we use that

$$b \triangleq \max_k \max_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}_k(x)} |\tilde{\imath}_k(x, y)| \quad (219)$$

is finite. Here, $\mathcal{Y}_k(x)$ denotes the support of $W_k(\cdot|x)$. We shall first treat the first K intervals and shortly return to the interval $\mathcal{T}_K$ for which we need the additional property that $\min_k I_k(P) \leq C$ every $P \in \mathcal{P}(\mathcal{X})$. We first obtain the following large-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\mathbb{P}\left[\max_{0\le n\le t}\tilde{\imath}_k(x^n;Y_k^n)\ge\tilde{\lambda}\right]=\mathbb{P}\left[\bigcup_{n=0}^{t}\left\{\tilde{\imath}_k(x^n;Y_k^n)\ge\tilde{\lambda}\right\}\right] \tag{204}$$

$$\le\mathbb{P}\left[\left\{\tilde{\imath}_k(x^t;Y_k^t)\ge\tilde{\lambda}\right\}\cup\bigcup_{n=0}^{t-1}\left(\left\{\tilde{\imath}_k(x^t;Y_k^t)\ge\tilde{\lambda}-\nu\right\}\cup\left\{\tilde{\imath}_k(x_{n+1}^t;Y_{k,n+1}^t)\le-\nu\right\}\right)\right] \tag{205}$$

$$=\mathbb{P}\left[\left\{\tilde{\imath}_k(x^t;Y_k^t)\ge\tilde{\lambda}-\nu\right\}\cup\bigcup_{n=1}^{t}\left\{\tilde{\imath}_k(x_n^t;Y_{k,n}^t)\le-\nu\right\}\right] \tag{206}$$

$$\le\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\ge\tilde{\lambda}-\nu\right]+\mathbb{P}\left[\bigcup_{n=1}^{t}\left\{\tilde{\imath}_k(x_n^t;Y_{k,n}^t)\le-\nu\right\}\right]. \tag{207}$$

deviation bound, which holds for all $t\in\mathcal{T}_i$, $i\in\{0,\cdots,K-1\}$ and $k\in\{i+1,\cdots,K\}$:

$$\max_{x^t\in\mathcal{X}^t}\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\ge\lambda\right]$$

$$=\max_{x^t\in\mathcal{X}^t}\mathbb{P}\left[\frac{\tilde{\imath}_k(x^t;Y_k^t)}{t}-I_k(P_{x^t})\ge\frac{\lambda}{t}-I_k(P_{x^t})\right] \tag{220}$$

$$\le\max_{x^t\in\mathcal{X}^t}\exp\left(-2t^2\frac{(\lambda/t-I_k(P_{x^t}))^2}{4tb^2}\right) \tag{221}$$

$$=\max_{x^t\in\mathcal{X}^t}\exp\left(-\frac{1}{2b^2}\left(\frac{\lambda-tI_k(P_{x^t})}{\sqrt{t}}\right)^2\right) \tag{222}$$

$$\le\max_{x^t\in\mathcal{X}^t}\exp\left(-\frac{1}{2b^2}\left(\frac{\lambda-t_{i+1}I_k(P_{x^t})}{\sqrt{t_{i+1}}}\right)^2\right) \tag{223}$$

$$\le\exp\left(-\frac{1}{2b^2}\left(\frac{\lambda-t_{i+1}C_{i+1}}{\sqrt{t_{i+1}}}\right)^2\right) \tag{224}$$

$$\le\exp\left(-\frac{1}{2b^2}\left(\frac{C_{i+1}\sqrt{V\lambda/C^3}\log\lambda}{\sqrt{\lambda/C_{i+1}-\sqrt{V\lambda/C^3}\log\lambda}}\right)^2\right) \tag{225}$$

$$\le\exp\left(-\frac{1}{2b^2}\left(\frac{C_{i+1}\sqrt{V\lambda/C^3}\log\lambda}{\sqrt{\lambda/C_{i+1}}}\right)^2\right) \tag{226}$$

$$=\exp\left(-\frac{1}{2b^2}\left(C_{i+1}^{3/2}\sqrt{V/C^3}\log\lambda\right)^2\right) \tag{227}$$

$$=\exp\Big(-\underbrace{\frac{VC_{i+1}^3}{2b^2C^3}}_{\triangleq c_{1i}}\log^2\lambda\Big) \tag{228}$$

$$=\left(\frac{1}{\lambda}\right)^{c_{1i}\log\lambda}. \tag{229}$$

Here, (221) follows from Hoeffding's inequality [24, Th. 2] and from (219), (223) follows because $(\lambda-tI_k(P_{x^t}))/\sqrt{t}$, for a fixed distribution P_{x^t}, is a nonincreasing function of t and because $t<t_{i+1}$, (224) follows because $I_k(P_{x^t})$ is uniformly upper-bounded by C_{i+1} for $k\in\{i+1,\cdots,K\}$ (recall that we assumed $C_1\ge\cdots\ge C_K$), (225) follows from (166), and (226)–(229) follow from algebraic manipulations.

Next, we consider the interval $\mathcal{T}_K$. Fix a probability distribution $P\in\mathcal{P}(\mathcal{X})$, and let

$$\kappa(P)\triangleq\arg\min_k I_k(P). \tag{230}$$

Note that $I_{\kappa(P)}(P)\le C$ for every $P\in\mathcal{P}(\mathcal{X})$. Let also $c_{1K}\triangleq V/(2b^2)$. We have the following bound for all $t\in\mathcal{T}_K$

$$\max_{x^t\in\mathcal{X}^t}\min_k\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\ge\lambda\right]$$

$$\le\max_{x^t\in\mathcal{X}^t}\min_k\exp\left(-\frac{1}{2b^2}\left(\frac{\lambda-t_{K+1}I_k(P_{x^t})}{\sqrt{t_{K+1}}}\right)^2\right) \tag{231}$$

$$\le\max_{x^t\in\mathcal{X}^t}\exp\left(-\frac{1}{2b^2}\left(\frac{\lambda-t_{K+1}I_{\kappa(P_{x^t})}(P_{x^t})}{\sqrt{t_{K+1}}}\right)^2\right) \tag{232}$$

$$\le\max_{x^t\in\mathcal{X}^t}\exp\left(-\frac{1}{2b^2}\left(\frac{\lambda-t_{K+1}C}{\sqrt{t_{K+1}}}\right)^2\right) \tag{233}$$

$$\le\left(\frac{1}{\lambda}\right)^{c_{1K}\log\lambda}. \tag{234}$$

Here, (231) follows from steps similar to (220)–(223), (233) holds because $I_{\kappa(P)}(P)\le C$ for every $P\in\mathcal{P}(\mathcal{X})$, and (234) follows from steps similar to (224)–(229). Using (229), we conclude that for all $i\in\{0,\cdots,K-1\}$

$$\sum_{t=t_i}^{t_{i+1}-1}L_t(\boldsymbol{\varepsilon})$$

$$\le\sum_{t=t_i}^{t_{i+1}-1}\max_{x^t\in\mathcal{X}^t}\prod_k\min\left\{1,\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\ge\lambda\right]+\varepsilon_k\right\}+\mathcal{O}(1) \tag{235}$$

$$\le\sum_{t=t_i}^{t_{i+1}-1}\max_{x^t\in\mathcal{X}^t}\prod_{k=i+1}^{K}\left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\ge\lambda\right]+\varepsilon_k\right)+\mathcal{O}(1) \tag{236}$$

$$\le\sum_{t=t_i}^{t_{i+1}-1}\prod_{k=i+1}^{K}\left(\lambda^{-c_{1i}\log\lambda}+\varepsilon_k\right)+\mathcal{O}(1) \tag{237}$$

$$\le(t_{i+1}-t_i)\prod_{k\in\{i+1,\cdots,K\}}\varepsilon_k+\mathcal{O}(1) \tag{238}$$

as $\lambda\to\infty$. Here, (238) follows because $(t_{i+1}-t_i)\lambda^{-c_{1i}\log\lambda}\le\mathfrak{c}\lambda^{-c_{1i}\log\lambda+1}=o(1)$ as $\lambda\to\infty$. Similarly, it follows from (234) that

$$\sum_{t=t_K}^{t_{K+1}-1}L_t(\boldsymbol{\varepsilon})$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\leq \sum_{t=t_K}^{t_{K+1}-1} \max_{x^t\in\mathcal{X}^t} \prod_k \min\left\{1, \mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right\} + \mathcal{O}(1) \quad (239)$$

$$\leq \sum_{t=t_K}^{t_{K+1}-1} \max_{x^t\in\mathcal{X}^t} \min_k \left\{\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right\} + \mathcal{O}(1) \quad (240)$$

$$\leq \sum_{t=t_K}^{t_{K+1}-1} \left(\max_{x^t\in\mathcal{X}^t} \min_k \mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\max_k\varepsilon_k\right) + \mathcal{O}(1) \quad (241)$$

$$\leq (t_{K+1}-t_K)\left(\lambda^{-c_{1K}\log\lambda}+\max_k\varepsilon_k\right)+\mathcal{O}(1) \quad (242)$$

$$\leq (t_{K+1}-t_K)\max_k\varepsilon_k+\mathcal{O}(1) \quad (243)$$

as $\lambda\to\infty$. Here, (240) follows because $\prod_k \min\{1,a_k\}\leq \min_k a_k$ for all nonnegative constants $\{a_k\}$, (241) holds because $\min_k\{a_k+\varepsilon_k\}\leq \min_k a_k+\max_k\varepsilon_k$, and (242) follows from (234). Finally, by adding (238) and (243), we obtain (244)–(247), shown in the top of the next page. Here, (246) follows by adding to (245) the term $(\max_k\varepsilon_k-\prod_k\varepsilon_k)\left(\lambda\rho/C_1-\sqrt{\lambda V/C^3}\log\lambda\right)$, which is positive for all sufficiently large λ; and (247) follows by the definition of the $\{d_i\}_{i\in\{0,\cdots,K\}}$ (see (169)–(171)). Finally, (175) follows from (247) and by the definition of $f(\boldsymbol{\varepsilon})$ in (172).

C. Central-limit analysis

Within the interval $[t_{K+1},\beta_{\boldsymbol{\varepsilon}}]$, we use Chebyshev's inequality and the Berry-Esseen central limit theorem to obtain a bound on (161). Fix a constant $\delta_2\in(0,1)$ and let $\mathcal{A}$ be a compact convex neighborhood of P^* such that for all $P\in\mathcal{A}$, we have both $V_k(P)>0$ and

$$\left|\sqrt{\frac{V_k(P)}{I_k(P)^3}}-\sqrt{\frac{V_k}{C^3}}\right|\leq\sqrt{\frac{V_k}{C^3}}\delta_2. \quad (248)$$

The existence of such a set $\mathcal{A}$ follows from the continuity of $I_k(\cdot)$ and $V_k(\cdot)$ at P^*. By the definition of $\mathcal{A}$, it follows that

$$I_{\kappa(P)}(P)<C-\delta_3 \quad (249)$$

for every $P\notin\mathcal{A}$ and for some $C>\delta_3>0$. This is a consequence of the uniqueness of P^*. The inequality (248) enables us to approximate $\sqrt{V_k(P)/I_k(P)^3}$ by $\sqrt{V_k/C^3}$ (which does not depend on P) as long as P belongs to the set $\mathcal{A}$. In other words, we can eliminate the dependency on P by introducing an error term proportional to δ_2, which can be made arbitrarily small.

We shall use the following upper bound on $L_t(\boldsymbol{\varepsilon})$:

$$L_t(\boldsymbol{\varepsilon}) \leq \max_{x^t\in\mathcal{X}^t}\prod_k\min\left\{1,\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right\} + \frac{2^K-1}{\lambda} \quad (250)$$

$$\leq \max\Bigg\{\max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\notin\mathcal{A}}}\min_k\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\max_k\varepsilon_k, \max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\in\mathcal{A}}}\prod_k\left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right)\Bigg\} + \frac{2^K-1}{\lambda}. \quad (251)$$

Here, (251) follows because $\prod_k(a_k+b_k)\leq\min_k(a_k+b_k)\leq\min_k a_k+\max_k b_k$ for all constants $\{a_k\}$ and $\{b_k\}$.

For all $x^t\in\mathcal{X}^t$ for which $P_{x^t}\notin\mathcal{A}$, we use Chebyshev's inequality to obtain the estimate

$$\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right] \leq \begin{cases}\frac{tV_k(P_{x^t})}{(\lambda-tI_k(P_{x^t}))^2} & \text{if } \lambda>tI_k(P_{x^t})\\ 1 & \text{otherwise}\end{cases}. \quad (252)$$

It follows from (249) and from the condition $t\leq\beta_{\boldsymbol{\varepsilon}}$ (see definition of $\beta_{\boldsymbol{\varepsilon}}$ in (164)) that $\lambda>tI_{\kappa(P)}(P)$ for every $P\notin\mathcal{A}$ and for all sufficiently large λ (recall the definition of $\kappa(\cdot)$ in (230)). Using (252), we obtain the following upper bound on the first term on the right-hand side of (251), which holds for $t\in[t_{K+1},\beta_{\boldsymbol{\varepsilon}}]$ and for all sufficiently large λ:

$$\max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\notin\mathcal{A}}}\min_k\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right] \leq \max_{P\notin\mathcal{A}}\frac{tV_{\kappa(P)}(P)}{(\lambda-tI_{\kappa(P)}(P))^2} \quad (253)$$

$$\leq \frac{V_{\max}t}{(\lambda-tC+t\delta_3)^2} \quad (254)$$

$$\leq \frac{2V_{\max}\lambda}{(\lambda\delta_3/C-\mathbb{c}\sqrt{\lambda}\log\lambda-\mathbb{c})^2} \quad (255)$$

$$\leq \frac{1}{\lambda}\underbrace{\frac{4V_{\max}C^2}{\delta_3}}_{\triangleq c_2}. \quad (256)$$

Here, (254) follows because there exists a constant $V_{\max}>0$ such that $V_k(P)\leq V_{\max}$ for every $P\in\mathcal{P}(\mathcal{X})$ [25, p. 7048] and because of (249); (255) follows because $t\leq\beta_{\boldsymbol{\varepsilon}}$ and $\delta_3<C$, which imply that, for all sufficiently large λ,

$$\lambda-tC+t\delta_3\geq\lambda-(C-\delta_3)\beta_{\boldsymbol{\varepsilon}} \quad (257)$$

$$=\lambda-(C-\delta_3)(\lambda/C+\mathbb{c}\sqrt{\lambda}\log\lambda) \quad (258)$$

$$=\lambda\delta_3/C-\mathbb{c}\sqrt{\lambda}\log\lambda-\mathbb{c}>0. \quad (259)$$

Finally, (256) holds for all sufficiently large λ. We see that (256) can be made arbitrarily close to zero by choosing λ sufficiently large. Now, we continue the chain of inequalities in (251) as follows:

$$L_t(\boldsymbol{\varepsilon}) \leq \max\left\{\frac{c_2}{\lambda}+\max_k\varepsilon_k, \max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\in\mathcal{A}}}\prod_k\left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right)\right\} + \frac{2^K-1}{\lambda} \quad (260)$$

$$\leq \max\left\{\frac{c_2}{\lambda}, \max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\in\mathcal{A}}}\prod_k\left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right)-\prod_k\varepsilon_k\right\} + \max_k\varepsilon_k+\frac{2^K-1}{\lambda} \quad (261)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\sum_{t=0}^{t_{K+1}-1} L_t(\boldsymbol{\varepsilon})$$

$$\leq \sum_{i=1}^{K}\left[(t_i - t_{i-1})\prod_{k\in\{i,\cdots,K\}}\varepsilon_k\right] + (t_{K+1}-t_K)\max_k \varepsilon_k + \mathcal{O}(1) \tag{244}$$

$$= \left(\frac{\lambda}{C_1} - \sqrt{\frac{\lambda V}{C^3}}\log\lambda\right)\prod_k \varepsilon_k + \sum_{i=2}^{K}\left[\left(\frac{\lambda}{C_i} - \frac{\lambda}{C_{i-1}}\right)\prod_{k\in\{i,\cdots,K\}}\varepsilon_k\right] + \left(\frac{\lambda}{C} - \frac{\lambda}{C_K}\right)\max_k \varepsilon_k + \mathcal{O}(1) \tag{245}$$

$$\leq \left(\frac{\lambda}{C_1} - \frac{\lambda\rho}{C_1}\right)\prod_k \varepsilon_k + \sum_{i=2}^{K}\left[\left(\frac{\lambda}{C_i} - \frac{\lambda}{C_{i-1}}\right)\prod_{k\in\{i,\cdots,K\}}\varepsilon_k\right] + \left(\frac{\lambda}{C} - \frac{\lambda}{C_K} + \frac{\lambda\rho}{C_1} - \sqrt{\frac{\lambda V}{C^3}}\log\lambda\right)\max_k \varepsilon_k + \mathcal{O}(1) \tag{246}$$

$$= \sum_{i=1}^{K}\left[d_i \prod_{k\in\{i,\cdots,K\}}\varepsilon_k\right] + \left(d_0 - \sqrt{\frac{\lambda V}{C^3}}\log\lambda\right)\max_k \varepsilon_k + \mathcal{O}(1). \tag{247}$$

$$\leq \max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\in\mathcal{A}}} \prod_k \left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right] + \varepsilon_k\right) - \prod_k \varepsilon_k + \max_k \varepsilon_k + \frac{1}{\lambda}\underbrace{\left(2^K - 1 + c_2\right)}_{\triangleq c_3}. \tag{262}$$

Here, in (261), we used that $\max_k \varepsilon_k \geq \prod_k \varepsilon_k$ because $\varepsilon_k \in [0,1]$ for all $k \in \mathcal{K}$. Note that the upper bound (262) allows us to consider only the $x^t \in \mathcal{X}^t$ for which $P_{x^t} \in \mathcal{P}(\mathcal{X})$. This in turn allows us to make use of (248). Specifically, for all $x^t \in \mathcal{X}^t$ for which $P_{x^t} \in \mathcal{A}$, the Berry-Esseen central limit theorem [26, Th. V.2.3] yields the following estimate:

$$\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]$$

$$\leq Q\left(\frac{\lambda - tI_k(P_{x^t})}{\sqrt{tV_k(P_{x^t})}}\right) + \frac{6tT_k(P_{x^t})}{(tV_k(P_{x^t}))^{3/2}} \tag{263}$$

$$\leq Q\left(\frac{\lambda - tI_k(P_{x^t})}{\sqrt{tV_k(P_{x^t})}}\right) + \frac{1}{\sqrt{2Ct}}\underbrace{\frac{6\sqrt{2C}\max_{P\in\mathcal{A}} T_k(P)}{\min_{P\in\mathcal{A}} V_k(P)^{3/2}}}_{\triangleq c_4} \tag{264}$$

$$\leq Q\left(\frac{\lambda/I_k(P_{x^t}) - t}{\sqrt{\lambda V_k(P_{x^t})/I_k(P_{x^t})^3}}\right) + \frac{c_4}{\sqrt{2Ct}} \tag{265}$$

$$\leq Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\lambda/I_k(P_{x^t}) - t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right) + \frac{c_4}{\sqrt{\lambda}}. \tag{266}$$

In (264), c_4 is a well-defined positive constant because $T_k(P) < \mathbb{c}$ uniformly [4, Lem. 46] and because the condition $V_k(P) > 0$ for $P \in \mathcal{A}$ combined with the compactness of $\mathcal{A}$ imply that $\min_{P\in\mathcal{A}} V_k(P)$ is well-defined and positive; (265) follows by the inequality (proven in Appendix VIII)

$$\frac{a - tb}{\sqrt{t}} \geq \frac{a - tb}{\sqrt{a/b}} \tag{267}$$

which holds for all positive a, b and t; and (266), which holds for all sufficiently large λ, follows from (248) (recall that $P_{x^t} \in \mathcal{A}$), which is equivalent to

$$\sqrt{\frac{V_k}{C^3}}(1-\delta_2) \leq \sqrt{\frac{V_k(P)}{I_k(P)^3}} \leq \sqrt{\frac{V_k}{C^3}}(1+\delta_2). \tag{268}$$

In (266), the role of ν_k is to select the upper or the lower bound in (268). The choice depends on the sign of $(\lambda/I_k(P_{x^t}) - t)$. The bound (266) implies that

$$\prod_k \left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right] + \varepsilon_k\right) \leq \prod_k \left(Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\lambda/I_k(P_{x^t}) - t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right) + \varepsilon_k\right) + \frac{c_5}{\sqrt{\lambda}} \tag{269}$$

where $c_5 \triangleq (2^K - 1)c_4$.

We shall next eliminate the dependency of the first term of the right-hand side of (269) on x^t by further upper-bounding this term. Let $P \in \mathcal{A}$; we have

$$Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\lambda/I_k(P) - t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right) \leq Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\lambda/(C + \nabla I_k(P-P^*)) - t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right) \tag{270}$$

$$\leq Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\frac{\lambda}{C} - \frac{\lambda}{C^2}\nabla I_k(P-P^*) - t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right). \tag{271}$$

Here, (270) follows because $I_k(P)$ is concave in P and because the Q function is monotonically decreasing; (271) follows from the inequality $\frac{a}{b+c} \geq \frac{a}{b} - \frac{a}{b^2}c$ which holds for all $a > 0, b > 0$, and $b + c > 0$. Indeed, $C + \nabla I_k(P - P^*) > 0$ for sufficiently small δ_2 since $P \in \mathcal{A}$. Using (271) in (269), we obtain the steps (272)–(274), shown in the top of the next page. Here, in (274), we used that $\varrho_k = \sqrt{V_k/V}$. Let now $\{H_{\delta_2,k}\}$ be i.i.d. RVs with cumulative distribution function

$$F_{H_{\delta_2,k}}(w) \triangleq Q\left(\min_{\nu_k\in\{-1,1\}} \frac{-w - \nabla I_k(\hat{\mathbf{v}}_{\delta_2}(w))}{\varrho_k(1+\delta_2\nu_k)}\right) \tag{275}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\prod_k \left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right) \leq \prod_k \left(Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\frac{\lambda}{C}-\frac{\lambda}{C^2}\nabla I_k(P_{x^t}-P^*)-t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right)+\varepsilon_k\right)+\frac{c_5}{\sqrt{\lambda}} \tag{272}$$

$$\leq \max_{P\in\mathcal{A}} \prod_k \left(Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\frac{\lambda}{C}-\frac{\lambda}{C^2}\nabla I_k(P-P^*)-t}{\sqrt{\lambda V_k/C^3}\,(1+\delta_2\nu_k)}\right)+\varepsilon_k\right)+\frac{c_5}{\sqrt{\lambda}} \tag{273}$$

$$\leq \max_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \prod_k \left(Q\left(\min_{\nu_k\in\{-1,1\}} \frac{\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}-\nabla I_k(\mathbf{v})}{\varrho_k(1+\delta_2\nu_k)}\right)+\varepsilon_k\right)+\frac{c_5}{\sqrt{\lambda}}. \tag{274}$$

---

where[9]

$$\hat{\mathbf{v}}_{\delta_2}(w) \triangleq \arg\max_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \prod_k \left(Q\left(\min_{\nu_k\in\{-1,1\}} \frac{-w-\nabla I_k(\mathbf{v})}{\varrho_k(1+\delta_2\nu_k)}\right)+\varepsilon_k\right). \tag{276}$$

We also denote by

$$\mathscr{K} \triangleq \{\tilde{\mathcal{K}}:\tilde{\mathcal{K}}\subseteq\mathcal{K}\}\setminus\{\emptyset\} \tag{277}$$

the set of all nonempty subsets of $\mathcal{K}$. Using these definitions and (274), we have that for every $x^t\in\mathcal{X}^t$ satisfying $P_{x^t}\in\mathcal{A}$,[10]

$$\begin{aligned}
L_t(\boldsymbol{\varepsilon}) &\leq \max_{\substack{x^t\in\mathcal{X}^t:\\ P_{x^t}\in\mathcal{A}}} \prod_k \left(\mathbb{P}\left[\tilde{\imath}_k(x^t;Y_k^t)\geq\lambda\right]+\varepsilon_k\right)-\prod_k\varepsilon_k \\
&\quad +\max_k\varepsilon_k+\frac{c_3}{\lambda} && (278)\\
&\leq \prod_k\left(F_{H_{\delta_2,k}}\left(-\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right)+\varepsilon_k\right)-\prod_k\varepsilon_k \\
&\quad +\max_k\varepsilon_k+\underbrace{2c_5}_{\triangleq c_6}\frac{1}{\sqrt{\lambda}} && (279)\\
&= \sum_{\bar{\mathcal{K}}\in\mathscr{K}}\left(\prod_{k\in\mathcal{K}\setminus\{\bar{\mathcal{K}}\}}\varepsilon_k\right)\left(\prod_{k\in\bar{\mathcal{K}}}F_{H_{\delta_2,k}}\left(-\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right)\right) \\
&\quad +\max_k\varepsilon_k+\frac{c_6}{\sqrt{\lambda}}. && (280)
\end{aligned}$$

Here, (279), which holds for sufficiently large λ, follows from (274) and (275); (280) follows from (277) and by expanding the product in the first term of (279) into 2^K terms.

We now evaluate (280). For every nonempty subset $\bar{\mathcal{K}}\subseteq\mathcal{K}$ and for sufficiently large λ, we have

$$\begin{aligned}
&\sum_{t=t_{K+1}}^{\beta_{\boldsymbol{\varepsilon}}}\prod_{k\in\bar{\mathcal{K}}}F_{H_{\delta_2,k}}\left(-\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right) \\
&\leq \int_{-\infty}^{\beta_{\boldsymbol{\varepsilon}}}\prod_{k\in\bar{\mathcal{K}}}F_{H_{\delta_2,k}}\left(-\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right)\mathrm{d}t+1 && (281)\\
&= \int_{-\infty}^{\beta_{\boldsymbol{\varepsilon}}}\mathbb{P}\left[\max_{k\in\bar{\mathcal{K}}}\left\{H_{\delta_2,k}+\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right\}\leq 0\right]\mathrm{d}t+1 && (282)\\
&= \int_{-\infty}^{\beta_{\boldsymbol{\varepsilon}}}\mathbb{P}\left[\max_{k\in\bar{\mathcal{K}}}\left\{\frac{\lambda}{C}+\sqrt{\frac{\lambda V}{C^3}}H_{\delta_2,k}\right\}\leq t\right]\mathrm{d}t+1 && (283)\\
&= \beta_{\boldsymbol{\varepsilon}}-\mathbb{E}\left[\min\left\{\beta_{\boldsymbol{\varepsilon}},\max_{k\in\bar{\mathcal{K}}}\left\{\frac{\lambda}{C}+\sqrt{\frac{\lambda V}{C^3}}H_{\delta_2,k}\right\}\right\}\right]+1 && (284)\\
&= \sqrt{\frac{\lambda V}{C^3}}\left(\nu_{\boldsymbol{\varepsilon}}-\mathbb{E}\left[\min\left\{\nu_{\boldsymbol{\varepsilon}},\max_{k\in\bar{\mathcal{K}}}H_{\delta_2,k}\right\}\right]\right)+1. && (285)
\end{aligned}$$

In (281), we have used that $F_{H_{\delta_2,k}}(\cdot)$ is a monotonically increasing function upper-bounded by one; (284) holds because for every RV X,

$$\mathbb{E}[\min\{a,X\}]=a-\int_{-\infty}^{a}\mathbb{P}[X\leq t]\,\mathrm{d}t. \tag{286}$$

Finally, (285) follows from (164).

Next, we substitute (285) into (280) and obtain a lower bound on $\sum_{t=t_{K+1}}^{\beta_{\boldsymbol{\varepsilon}}}(1-L_t(\boldsymbol{\varepsilon}))$ through the steps (287)–(290), shown in the top of the next page. This lower bound holds for all sufficiently large λ and for all $\boldsymbol{\varepsilon}\in[0,1]^K$ and the function $g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon})$ is defined in (291), shown in the top of the next page. In (287), we have used (280) and (288) follows by interchanging the order of summations. We also used that $\nu_{\boldsymbol{\varepsilon}}$ is bounded from above, which implies that there exists a constant c_7 that does not depend on $\boldsymbol{\varepsilon}\in[0,1]^K$ such that $\sqrt{V/C^3}c_6(\nu_{\boldsymbol{\varepsilon}}+\log\lambda)+2^K\leq c_7\log\lambda$ for all sufficiently large λ (recall that $|\mathscr{K}|=2^K-1$). Finally, (289)follows from (285). This establishes (176).

D. Optimization over $f(\cdot)$

We observe that (168) implies that, for every $i\in\{1,\cdots,K-1\}$,

$$\sum_{j=1}^{i}(d_j-d_0)>0. \tag{292}$$

In turn, these inequalities imply that there exist constants $\{\zeta_i\}_{i\in\{1,\cdots,K-2\}}$ such that

$$d_i-d_0+\zeta_{i-1}-\zeta_i>0 \tag{293}$$

for every $i\in\{1,\cdots,K-1\}$. Here, we have set $\zeta_0\triangleq\zeta_{K-1}\triangleq 0$ for convenience. A proof of this claim can be found in Lemma 14 in Appendix VIII.

We use (293) to show that $f(\boldsymbol{\varepsilon})$ is lower-bounded by an affine function through the steps (294)–(297), shown in the top of the next page. Here, (294) holds because $\prod_{k\in\{i,\cdots,K\}}\varepsilon_k\leq$

[9]If the maximizer of (276) is not unique, $\hat{\mathbf{v}}_{\delta_2}(w)$ is chosen arbitrarily from the set of maximizers.

[10]In (280), we use the convention that $\prod_{k\in\emptyset}a_k=1$ for every $a_k\in\mathbb{R}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\sum_{t=t_{K+1}}^{\beta_{\boldsymbol{\varepsilon}}} (1-L_t(\boldsymbol{\varepsilon}))$$

$$\geq \sqrt{\frac{\lambda V}{C^3}}(\log\lambda+\nu_{\boldsymbol{\varepsilon}})\left(1-\max_k \varepsilon_k\right) - \sum_{t=t_{K+1}}^{\beta_{\boldsymbol{\varepsilon}}}\left[\sum_{\bar{\mathcal{K}}\in\mathscr{K}}\left(\prod_{k\in\mathcal{K}\setminus\{\bar{\mathcal{K}}\}}\varepsilon_k\right)\left(\prod_{k\in\bar{\mathcal{K}}} F_{H_{\delta_2,k}}\left(-\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right)\right)\right] - 2^K - \sqrt{\frac{V}{C^3}}c_6(\log\lambda+\nu_{\boldsymbol{\varepsilon}}) \tag{287}$$

$$= \sqrt{\frac{\lambda V}{C^3}}(\log\lambda+\nu_{\boldsymbol{\varepsilon}})\left(1-\max_k \varepsilon_k\right) - \sum_{\bar{\mathcal{K}}\in\mathscr{K}}\left(\prod_{k\in\mathcal{K}\setminus\{\bar{\mathcal{K}}\}}\varepsilon_k\right)\left(\sum_{t=t_{K+1}}^{\beta_{\boldsymbol{\varepsilon}}}\prod_{k\in\bar{\mathcal{K}}} F_{H_{\delta_2,k}}\left(-\frac{\lambda/C-t}{\sqrt{\lambda V/C^3}}\right)\right) - c_7\log\lambda \tag{288}$$

$$\geq \sqrt{\frac{\lambda V}{C^3}}\left((\log\lambda+\nu_{\boldsymbol{\varepsilon}})(1-\max_k\varepsilon_k) - \sum_{\bar{\mathcal{K}}\in\mathscr{K}}\left(\prod_{k\in\mathcal{K}\setminus\{\bar{\mathcal{K}}\}}\varepsilon_k\right)\left(\nu_{\boldsymbol{\varepsilon}} - \mathbb{E}\Big[\min\Big\{\nu_{\boldsymbol{\varepsilon}}, \max_{k\in\bar{\mathcal{K}}} H_{\delta_2,k}\Big\}\Big]\right)\right) - c_7\log\lambda \tag{289}$$

$$\geq \sqrt{\lambda}g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}) + \sqrt{\frac{\lambda V}{C^3}}(1-\max_k\varepsilon_k)\log\lambda - c_7\log\lambda. \tag{290}$$

$$g_{\delta_1,\delta_2}(\boldsymbol{\varepsilon}) \triangleq \sqrt{\frac{V}{C^3}}\left(\nu_{\boldsymbol{\varepsilon}}(1-\max_k\varepsilon_k) - \sum_{\bar{\mathcal{K}}\in\mathscr{K}}\left(\prod_{k\in\mathcal{K}\setminus\{\bar{\mathcal{K}}\}}\varepsilon_k\right)\left(\nu_{\boldsymbol{\varepsilon}} - \mathbb{E}\Big[\min\Big\{\nu_{\boldsymbol{\varepsilon}}, \max_{k\in\bar{\mathcal{K}}} H_{\delta_2,k}\Big\}\Big]\right)\right). \tag{291}$$

$$f(\boldsymbol{\varepsilon}) \geq \frac{1}{C} - \sum_{i=1}^{K} d_i \min_{k\in\{i,\cdots,K\}}\varepsilon_k - d_0\max_k\varepsilon_k \tag{294}$$

$$= \frac{1}{C} - \sum_{i=1}^{K-1}(d_i-d_0)\min_{k\in\{i,\cdots,K\}}\varepsilon_k - d_0\left(\max_k\varepsilon_k + \sum_{i=1}^{K-1}\min_{k\in\{i,\cdots,K\}}\varepsilon_k\right) - d_K\varepsilon_K \tag{295}$$

$$\geq \frac{1}{C} - \sum_{i=1}^{K-1}(d_i-d_0+\zeta_{i-1}-\zeta_i)\min_{k\in\{i,\cdots,K\}}\varepsilon_k - d_0(\varepsilon_1+\cdots+\varepsilon_K) - d_K\varepsilon_K \tag{296}$$

$$\geq \frac{1}{C} - \sum_{i=1}^{K-1}(d_i-d_0+\zeta_{i-1}-\zeta_i)\frac{\varepsilon_i+\cdots+\varepsilon_K}{K-i+1} - d_0(\varepsilon_1+\cdots+\varepsilon_K) - d_K\varepsilon_K. \tag{297}$$

$\min_{k\in\{i,\cdots,K\}}\varepsilon_k$; in (296), we used that $\min_{k\in\{i,\cdots,K\}}\varepsilon_k \leq \min_{k\in\{i+1,\cdots,K\}}\varepsilon_k$ for $i\in\{1,\cdots,K-1\}$, which implies that

$$\sum_{i=1}^{K-1}(\zeta_{i-1}-\zeta_i)\min_{k\in\{i,\cdots,K\}}\varepsilon_k$$

$$\geq \sum_{i=1}^{K-1}\zeta_{i-1}\min_{k\in\{i,\cdots,K\}}\varepsilon_k - \sum_{i=1}^{K-1}\zeta_i\min_{k\in\{i+1,\cdots,K\}}\varepsilon_k \tag{298}$$

$$= \sum_{i=1}^{K-1}\zeta_{i-1}\min_{k\in\{i,\cdots,K\}}\varepsilon_k - \sum_{i=2}^{K}\zeta_{i-1}\min_{k\in\{i,\cdots,K\}}\varepsilon_k \tag{299}$$

$$= \zeta_0\varepsilon_1 - \zeta_{K-1}\varepsilon_K = 0. \tag{300}$$

Furthermore, (297) holds because $\min_{k\in\{i,\cdots,K\}}\varepsilon_k \leq (\varepsilon_i+\cdots+\varepsilon_K)/(K-i+1)$. It follows from (293) that $(d_i-d_0+\zeta_{i-1}-\zeta_i)$ is positive for all $i\in\{1,\cdots,K-1\}$. Hence, the inequality (297) holds with equality if and only if $\varepsilon_1=\cdots=\varepsilon_K$. This implies that a necessary and sufficient condition for the chain of inequalities (294)–(297) to hold with equality is that $\varepsilon_1=\cdots=\varepsilon_K$ and that $\varepsilon_1\in\{0,1\}$. This implies that

$$\min_{\substack{P_U,\boldsymbol{\varepsilon}^{(u)}\in[0,1]^K:\\ \mathbb{E}_U\left[\varepsilon_k^{(U)}\right]\leq\epsilon+(\log M)^{-1}}} \mathbb{E}_U\left[f(\boldsymbol{\varepsilon}^{(U)})\right]$$

$$= (1-\epsilon-(\log M)^{-1})f(\mathbf{0}_K) + (\epsilon+(\log M)^{-1})f(\mathbf{1}_K) \tag{301}$$

which is equivalent to (183).

Appendix V
Proof of Theorem 5 (achievability) and of Theorem 8

First, we prove Theorem 8. Then, we show that the achievability part of Theorem 5 follows as a special case of Theorem 8. To establish Theorem 8, we make use of the following lemma.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Lemma 12: Under the conditions of Theorem 8, there exists a joint probability distribution P_{X^∞} on $\mathcal{X}^\infty$ such that the stopping times $\{\tau_k(\gamma)\}$

$$\tau_k(\gamma) \triangleq \inf\big\{n \geq 0 : i_{P_{X^n},W_k^n}(X^n;Y_k^n) \geq \gamma\big\} \quad (302)$$

satisfy

$$\mathbb{E}\Big[\max_k \tau_k(\gamma)\Big] \leq \frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\mathbb{E}\Big[\max_k \bar{H}_k\Big] + o(\sqrt{\gamma}). \quad (303)$$

Here, the independent RVs $\{\bar{H}_k\}$ have cumulative distribution functions given in (96) and P_{X^n} in (302) denotes the joint probability distribution of the first n entries of $X^\infty \sim P_{X^\infty}$.

Proof: See Appendix VII. ∎

Define now the function

$$g(x) \triangleq \sqrt{\frac{xV}{C^3}}\mathbb{E}\Big[\max_k \bar{H}_k\Big] \quad (304)$$

and let P_{X^∞} be distributed according to Lemma 12. In view of Theorem 1, let for all integers $\bar{\ell} > 0$ and for all $\delta > 0$

$$\gamma_{\bar{\ell}} \triangleq C\left(\bar{\ell} - (1+\delta)g(C\bar{\ell})\right) \quad (305)$$

$$q_{\bar{\ell}} \triangleq \frac{\bar{\ell}\epsilon - 1}{\bar{\ell} - 1} \quad (306)$$

$$M_{\bar{\ell}} \triangleq \left\lfloor \exp\big(\gamma_{\bar{\ell}} - \log\bar{\ell}\big)\right\rfloor. \quad (307)$$

Then, we have (cf. (21))

$$q_{\bar{\ell}} + (1-q_{\bar{\ell}})(M_{\bar{\ell}}-1)\exp\{-\gamma_{\bar{\ell}}\} \leq \frac{\bar{\ell}\epsilon - 1}{\bar{\ell}-1} + \frac{\bar{\ell}(1-\epsilon)}{\bar{\ell}-1}\frac{1}{\bar{\ell}} = \epsilon. \quad (308)$$

Additionally, suppose that there exists an integer $\ell_0 \geq 0$ such that, for all $\bar{\ell} > \ell_0$,

$$\mathbb{E}\Big[\max_k \tau_k(\gamma_{\bar{\ell}})\Big] \leq \bar{\ell} \quad (309)$$

and $M_{\bar{\ell}} \geq 2$. Then, for $\bar{\ell} \geq \ell_0$, we have

$$(1-q_{\bar{\ell}})\mathbb{E}\Big[\max_k \tau_k(\gamma_{\bar{\ell}})\Big] \leq \frac{\bar{\ell}(1-\epsilon)}{\bar{\ell}-1}\bar{\ell} \triangleq \ell_{\bar{\ell}}. \quad (310)$$

By invoking Theorem 1 with $q = q_{\bar{\ell}}$, $\gamma = \gamma_{\bar{\ell}}$, and $M = M_{\bar{\ell}}$, and by using (21) along with the inequalities (308) and (310), we conclude that there exists a sequence of $(\ell_{\bar{\ell}}, M_{\bar{\ell}}, \epsilon)$-VLSF codes for all $\bar{\ell} \geq \ell_0$. Consequently, we have that for all $\bar{\ell} \geq \ell_0$,

$$\begin{aligned}
&\log M^*_{\text{sf}}(\ell_{\bar{\ell}}, \epsilon) \\
&\quad \geq \log M_{\bar{\ell}} && (311)\\
&\quad \geq C\left(\bar{\ell} - (1+\delta)g(C\bar{\ell})\right) - \log\bar{\ell} - 1 && (312)\\
&\quad = \frac{C\ell_{\bar{\ell}}}{1-\epsilon} - (1+\delta)\sqrt{\frac{V\ell_{\bar{\ell}}}{1-\epsilon}}\mathbb{E}\Big[\max_k \bar{H}_k\Big] + \mathcal{O}(1). && (313)
\end{aligned}$$

Here, in (312), we used that $\log(\lfloor x\rfloor) \geq \log(x-1) \geq \log x - 1$ for $x \geq 2$; furthermore, (313) follows because

$$\ell_{\bar{\ell}} = \frac{(\bar{\ell})^2(1-\epsilon)}{\bar{\ell}-1} \leq \bar{\ell}(1-\epsilon) + o(1). \quad (314)$$

Since we can choose δ arbitrarily small, we conclude that (313) implies (94).

Proof of (309)*:* By Lemma 12, there exists an integer ℓ_0 such that, for all $\bar{\ell} \geq \ell_0$, we have

$$\begin{aligned}
\mathbb{E}\Big[\max_k \tau_k(\gamma_{\bar{\ell}})\Big] &\leq \frac{\gamma_{\bar{\ell}}}{C} + (1+\delta)g(\gamma_{\bar{\ell}}) && (315)\\
&= \bar{\ell} - (1+\delta)g(C\bar{\ell}) \\
&\quad + (1+\delta)g(C\bar{\ell} - C(1+\delta)g(C\bar{\ell})) && (316)\\
&\leq \bar{\ell}. && (317)
\end{aligned}$$

Here, (316) follows by the definition of $\gamma_{\bar{\ell}}$ in (305), and (317) holds because $g(x)$ is nonnegative and nondecreasing, which implies that $g(C\bar{\ell}) - g(C\bar{\ell} - C(1+\delta)g(C\bar{\ell})) \geq 0$.

Proof of the achievability part of Theorem 5: The achievability bound in Theorem 5 follows by setting $\bar{\mathbf{v}}(w)$ in Theorem 8 equal to the constant vector

$$\bar{\mathbf{v}}_{\text{const}} \triangleq -\arg\min_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \mathbb{E}\left[\max_k \nabla I_k(\mathbf{v}) + \varrho_k Z_k\right] \quad (318)$$

where $Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. This implies that $\bar{\mathbf{v}}'(w) = \mathbf{0}_{|\mathcal{X}|}$. Hence, $E_k(s) = 0$. In this case, we have that for every $w \in \mathbb{R}$,

$$\begin{aligned}
F_{\bar{H}_k}(w) &= \Phi\left(\frac{w + \nabla I_k(\bar{\mathbf{v}}_{\text{const}})}{\varrho_k}\right) && (319)\\
&= \mathbb{P}[-\nabla I_k(\bar{\mathbf{v}}_{\text{const}}) + \varrho_k Z_k \leq w]. && (320)
\end{aligned}$$

Hence, the $\{\bar{H}_k\}$ have the same distribution as $\{-\nabla I_k(\bar{\mathbf{v}}_{\text{const}}) + \varrho_k Z_k\}$. The achievability part of Theorem 5 is established by noting that

$$\begin{aligned}
\mathbb{E}\Big[\max_k \bar{H}_k\Big] &= \mathbb{E}\Big[\max_k\{-\nabla I_k(\bar{\mathbf{v}}_{\text{const}}) + \varrho_k Z_k\}\Big] && (321)\\
&= \min_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \mathbb{E}\Big[\max_k\{\nabla I_k(\mathbf{v}) + \varrho_k Z_k\}\Big]. && (322)
\end{aligned}$$

Appendix VI
Proof of Lemma 10

Our objective is to show that, if the conditions in Lemma 10 are satisfied, then $\boldsymbol{\beta}(w)$ given in (98) satisfies

$$P^*(x) + CP^*_{r(x)}(x)\beta'_{r(x)}(w) \in [0,1] \quad (323)$$

for every $x \in \mathcal{X}$ and every $w \in \mathbb{R}$. First, given $w \in \mathbb{R}$, we shall analyze the function

$$v(w) \triangleq \arg\max_{v\in\mathbb{R}} \prod_{k=1}^{2} \Phi\left(\frac{1}{\varrho_k}(w + v\Delta_k)\right). \quad (324)$$

The objective function in (324) is differentiable everywhere in v. Furthermore, it is strictly log-concave in v because $\Phi(\cdot)$ is strictly log-concave, and it tends to 0 as $|v| \to \infty$ (recall that $\Delta_1 > 0$ and $\Delta_2 < 0$). This implies that (324) has exactly one maximum which is also the unique stationary point.

By (98), we have that $\boldsymbol{\beta}(w) = [v(w), -v(w)]^{\mathrm{T}}$. Therefore (323) is equivalent to

$$P^*(x) + (-1)^{r(x)+1}CP^*_{r(x)}v'(w) \in [0,1] \quad (325)$$

for every $x \in \mathcal{X}$ and every $w \in \mathbb{R}$. We characterize $v'(w)$ in two steps. First, we show that $v'(w) > D$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$, that $v'(w) < D$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 < 0$, and that $v'(w) = D$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 = 0$. Next, we show that $v(\cdot)$ is monotonic.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Specifically, we demonstrate that (122) implies that $v'(w) < 0$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$ and $v'(w) > 0$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 < 0$. By combining the two steps, we find that $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$ implies $D < v'(w) < 0$ and that $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 < 0$ implies $0 < v'(w) < D$. This argument establishes (121) because $P^*(x) \in [0,1]$ for all $x \in \mathcal{X}$.

First step: Let $\psi(x) \triangleq \phi(x)/\Phi(x)$. We characterize the stationary point of the objective function in (324) by taking its logarithm, by differentiating it with respect to v, and by equating the resulting expression to zero:

$$\sum_{k=1}^{2} \psi\left(\frac{w + v\Delta_k}{\varrho_k}\right) \frac{\Delta_k}{\varrho_k} = 0. \tag{326}$$

One can readily verify that (326) has exactly one solution, which we denote by $v(w)$ to emphasize its dependence on w. Our objective is to characterize $v'(w)$. Hence, we differentiate both sides of (326) with respect to w to obtain an implicit equation for $v'(w)$:

$$\sum_{k=1}^{2} \psi'\left(\frac{w + v(w)\Delta_k}{\varrho_k}\right) \frac{\Delta_k}{\varrho_k^2} \left(1 + v'(w)\Delta_k\right) = 0. \tag{327}$$

Let $\bar{v}(w, v)$ be the solution to the following equation in z

$$\sum_{k=1}^{2} \psi'\left(\frac{w + v\Delta_k}{\varrho_k}\right) \frac{\Delta_k}{\varrho_k^2} \left(1 + z\Delta_k\right) = 0. \tag{328}$$

Note that we must have $v'(w) = \bar{v}(w, v(w))$. Solving (328) for z, we obtain that

$$\bar{v}(w, v) = -\frac{\displaystyle\sum_{k=1}^{2} \frac{\Delta_k}{\varrho_k^2} \psi'\left(\frac{w + v\Delta_k}{\varrho_k}\right)}{\displaystyle\sum_{k=1}^{2} \frac{\Delta_k^2}{\varrho_k^2} \psi'\left(\frac{w + v\Delta_k}{\varrho_k}\right)} \tag{329}$$

for every $w \in \mathbb{R}$ and every $v \in \mathbb{R}$. Since $\psi'(\cdot) \in (-1, 0)$, since $\psi'(x)$ is an increasing function in x (this result is proven in Lemma 15(a)-(b) in Appendix VIII), and since $\Delta_1 > 0 > \Delta_2$, we conclude that $\bar{v}(w, v)$ is an increasing function of v for fixed w.

We proceed by noting that the following equation in ζ

$$\frac{w + \zeta w\Delta_1}{\varrho_1} = \frac{w + \zeta w\Delta_2}{\varrho_2} \tag{330}$$

is solved by

$$\zeta \triangleq \frac{\varrho_1 - \varrho_2}{\Delta_1\varrho_2 - \Delta_2\varrho_1}. \tag{331}$$

For the case $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 = 0$, we observe that $\zeta = D$ (recall that D is defined in (120)) and that $v(w) = Dw$ solves (326). Next, consider the case $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$. Define

$$\kappa_w(a) \triangleq \sum_{k=1}^{2} \psi\left(\frac{w + a\Delta_k}{\varrho_k}\right) \frac{\Delta_k}{\varrho_k}. \tag{332}$$

Note that $\kappa_w(a)$ is a decreasing function in a because $\psi(\cdot)$ is a decreasing function (proved in Lemma 15(a)) and because $\Delta_1 > 0 > \Delta_2$. Additionally, (326) implies that $\kappa_w(v(w)) = 0$. Now, we use (330), the positivity of $\psi(\cdot)$, and that $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$, to conclude that

$$\begin{aligned}
\kappa_w(\zeta w) &= \sum_{k=1}^{2} \psi\left(\frac{w + \zeta w\Delta_k}{\varrho_k}\right) \frac{\Delta_k}{\varrho_k} && (333)\\
&= \psi\left(\frac{w + \zeta w\Delta_1}{\varrho_1}\right) \left(\frac{\Delta_1}{\varrho_1} + \frac{\Delta_2}{\varrho_2}\right) && (334)\\
&> 0 && (335)\\
&= \kappa_w(v(w)). && (336)
\end{aligned}$$

Since $\kappa_w(\cdot)$ is a decreasing function, (336) implies that $v(w) > \zeta w$. Thus, using that $\bar{v}(w, v)$ is increasing in v for fixed w, we conclude that

$$v'(w) = \bar{v}(w, v(w)) > \bar{v}(w, \zeta w) = D. \tag{337}$$

Following a similar line of reasoning, one can show that $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 < 0$ implies $v'(w) < D$.

Second step: Next, we show that $v(w)$ is monotonic. First, note that the solutions v^* and w^* of the optimization problem

$$\max_{v,w} \left\{ e^{-w\alpha} \prod_{k=1}^{2} \Phi\left(\frac{1}{\varrho_k}\left(w + v\Delta_k\right)\right) \right\} \tag{338}$$

where $\alpha > 0$, satisfy $v^* = v(w^*)$. The objective function in (338) is differentiable everywhere, strictly log-concave in w and v, and it tends to zero as $|v| \to \infty$ or $|w| \to \infty$. These properties imply that (338) has exactly one maximum, which is also the unique stationary point. We let $v^*(\alpha)$ and $w^*(\alpha)$ denote the solution of (338). Then, for each $\tilde{w} \in \mathbb{R}$, there exists a constant $\tilde{\alpha} > 0$ such that $v(\tilde{w}) = v^*(\tilde{\alpha})$ and $\tilde{w} = w^*(\tilde{\alpha})$. Hence, the set of points $\{(v(w), w)\}_{w\in\mathbb{R}}$ is equivalently parameterized by set of points $\{(v^*(\alpha), w^*(\alpha)\}_{\alpha>0}$. We also point out that $w^*(\alpha)$ is nonincreasing in α. By taking the logarithm of the objective function (338) and equating it to zero, we obtain the following stationarity conditions for $v^*(\alpha)$ and $w^*(\alpha)$

$$\begin{aligned}
\sum_{k=1}^{2} \psi\left(\frac{w^*(\alpha) + v^*(\alpha)\Delta_k}{\varrho_k}\right) \frac{1}{\varrho_k} &= \alpha && (339)\\
\sum_{k=1}^{2} \psi\left(\frac{w^*(\alpha) + v^*(\alpha)\Delta_k}{\varrho_k}\right) \frac{\Delta_k}{\varrho_k} &= 0. && (340)
\end{aligned}$$

Solving (339) and (340) for $v^*(\alpha)$ and $w^*(\alpha)$, we find that

$$\begin{aligned}
v^*(\alpha) = \frac{1}{\Delta_1 - \Delta_2} \Bigg(& \varrho_1 \psi^{-1}\left(\frac{\alpha\varrho_1}{1 - \Delta_1/\Delta_2}\right) \\
& - \varrho_2 \psi^{-1}\left(\frac{\alpha\varrho_2}{1 - \Delta_2/\Delta_1}\right) \Bigg) && (341)\\
w^*(\alpha) = \frac{\Delta_1\Delta_2}{\Delta_2 - \Delta_1} \Bigg(& \frac{\varrho_1}{\Delta_1} \psi^{-1}\left(\frac{\alpha\varrho_1}{1 - \Delta_1/\Delta_2}\right) \\
& - \frac{\varrho_2}{\Delta_2} \psi^{-1}\left(\frac{\alpha\varrho_2}{1 - \Delta_2/\Delta_1}\right) \Bigg). && (342)
\end{aligned}$$

Note that $\Delta_1 - \Delta_2 > 0$ due to the assumption $\Delta_1 > 0 > \Delta_2$. Since $w^*(\alpha)$ is a nonincreasing function of α, our objective is to show that $v^*(\alpha)$ is either nonincreasing or nondecreasing in α. In particular, if $v^*(\alpha)$ is nonincreasing then $v(w)$ must be nondecreasing and vice versa. Let $g(x) \triangleq -\psi'(\psi^{-1}(x))$ for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$x > 0$. By taking the derivative of (341) with respect to α, we obtain

$$\frac{\partial v^*}{\partial \alpha} = \frac{1}{\Delta_1 - \Delta_2}\left(\frac{\varrho_2^2/(1-\Delta_2/\Delta_1)}{g\Big(\frac{\alpha\varrho_2}{1-\Delta_2/\Delta_1}\Big)} - \frac{\varrho_1^2/(1-\Delta_1/\Delta_2)}{g\Big(\frac{\alpha\varrho_1}{1-\Delta_1/\Delta_2}\Big)}\right). \tag{343}$$

Note that $g(x)$ is a positive function, which satisfies the following property: $\beta g(x) \geq g(\beta x)$ for $\beta \geq 1$ and $x > 0$ (this is proved in Lemma 15(c) in Appendix VIII). This implies that, when $\beta > 1$,

$$\frac{g(\beta x)}{g(x)} < \beta. \tag{344}$$

Similarly, when $0 < \beta < 1$, one readily finds that (344) implies

$$\frac{g(\beta x)}{g(x)} > \frac{1}{\beta}. \tag{345}$$

To show monotonicity of $v^*(\alpha)$, we analyze the sign of (343). Let

$$h(\alpha) \triangleq \frac{\varrho_2^2(1-\Delta_1/\Delta_2)g\Big(\frac{\alpha\varrho_1}{1-\Delta_1/\Delta_2}\Big)}{\varrho_1^2(1-\Delta_2/\Delta_1)g\Big(\frac{\alpha\varrho_2}{1-\Delta_2/\Delta_1}\Big)} \tag{346}$$

$$= -\frac{\varrho_2^2\Delta_1 g\Big(\frac{\alpha\varrho_1}{1-\Delta_1/\Delta_2}\Big)}{\varrho_1^2\Delta_2 g\Big(\frac{\alpha\varrho_2}{1-\Delta_2/\Delta_1}\Big)}. \tag{347}$$

Note that $h(\alpha) > 1$ implies $\partial v^*/\partial\alpha > 0$. Furthermore, $h(\alpha) < 1$ implies $\partial v^*/\partial\alpha < 0$.

Consider the case $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$. By (122), we must have that $\varrho_2 \geq \varrho_1$. But this implies that

$$h(\alpha) > \frac{\varrho_2}{\varrho_1}\frac{g\Big(\frac{\alpha\varrho_1}{1-\Delta_1/\Delta_2}\Big)}{g\Big(\frac{\alpha\varrho_2}{1-\Delta_2/\Delta_1}\Big)} \tag{348}$$

$$\geq -\frac{\varrho_2}{\varrho_1}\frac{\varrho_2\Delta_1}{\varrho_1\Delta_2} \tag{349}$$

$$\geq \frac{\varrho_2}{\varrho_1} \tag{350}$$

$$\geq 1. \tag{351}$$

Here, (348) follows from $\Delta_1\varrho_2/(\Delta_2\varrho_1) < -1$; (349) follows from (345) with $x = \alpha\varrho_2/(1-\Delta_2/\Delta_1) > 0$ and $\beta = -\varrho_1\Delta_2/(\varrho_2\Delta_1) \in (0,1)$, (350) holds because $\Delta_1\varrho_2/(\Delta_2\varrho_1) < -1$, and (351) holds because $\varrho_2/\varrho_1 \geq 1$. Using a similar line of reasoning, we obtain for the case $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 < 0$ that

$$h(\alpha) < 1. \tag{352}$$

Using that $h(\alpha) > 1$ implies $\partial v^*/\partial\alpha > 0$ and that $h(\alpha) < 1$ implies $\partial v^*/\partial\alpha < 0$, we conclude that $D < v'(w) < 0$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 > 0$ and that $D > v'(w) > 0$ if $\Delta_1/\varrho_1 + \Delta_2/\varrho_2 < 0$.

Appendix VII
Proof of Lemma 12

To prove Lemma 12, we shall first construct a suitable non-stationary joint probability distribution P_{X^∞} on $\mathcal{X}^\infty$. Then we shall set

$$\beta_- \triangleq \left\lfloor \frac{\gamma}{C} - \sqrt{\frac{\gamma V}{C^3}}\log\gamma \right\rfloor \tag{353}$$

and

$$\beta_+ \triangleq \left\lfloor \frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\log\gamma \right\rfloor \tag{354}$$

and compute an asymptotic upper bound on

$$\mathbb{E}\Big[\max_k \tau_k\Big] = \sum_{t=0}^{\infty}\Big(1 - \mathbb{P}\Big[\max_k \tau_k \leq t\Big]\Big) \tag{355}$$

$$\leq \beta_- + 1 + \sum_{t=\beta_-+1}^{\beta_+}\Big(1 - \mathbb{P}\Big[\max_k \tau_k \leq t\Big]\Big) + \sum_{t=\beta_++1}^{\infty}\Big(1 - \mathbb{P}\Big[\max_k \tau_k \leq t\Big]\Big) \tag{356}$$

that matches (303) in the limit $\gamma \to \infty$. This is done by obtaining lower bounds on $\mathbb{P}[\max_k \tau_k \leq t]$ for $t \geq \beta_- + 1$. Similarly as in Appendix IV, it turns out convenient to treat the two subintervals $[\beta_- + 1, \beta_+]$ and $[\beta_+ + 1, \infty)$ differently. In the former subinterval, our main tool is a multivariate version of the Berry-Esseen central limit theorem for sums of independent RVs. In the latter subinterval, we apply Hoeffding's inequality. To compute the desired lower bound on $\mathbb{P}[\max_k \tau_k \leq t]$, we shall use the following relation between τ_k and $i_{P_{X^t},W_k^t}(X^t;Y_k^t)$, which follows from the definition of τ_k in (302):

$$\mathbb{P}\Big[\max_k \tau_k \leq t\Big] \geq \mathbb{P}\Big[\min_k i_{P_{X^t},W_k^t}(X^t;Y_k^t) \geq \gamma\Big]. \tag{357}$$

We next summarize the key steps of the proof. Details are provided in Appendix VII-A–VII-D.

Step 1: First, we specify the probability distribution P_{X^∞} on $\mathcal{X}^\infty$ for which (303) holds. Let

$$w(t) \triangleq \frac{t - \gamma/C}{\sqrt{\gamma V/C^3}} \tag{358}$$

$$\bar{P}^{(1)}(\gamma) \triangleq P^* + \sqrt{VC/\gamma}\,\bar{\mathbf{v}}(w(\beta_-)) \tag{359}$$

$$P^{(2)}(w) \triangleq P^* + C\bar{\mathbf{v}}'(w) \tag{360}$$

$$P^{(3)}(\gamma) \triangleq P^* + \sqrt{VC/\gamma}\,\bar{\mathbf{v}}(w(\beta_+)) \tag{361}$$

and let $P^{(1)}(\gamma) \in \mathcal{P}_{\beta_-}(\mathcal{X})$ be the type that minimizes $\big\|\bar{P}^{(1)}(\gamma) - P^{(1)}(\gamma)\big\|$ (recall that $\mathcal{P}_n(\mathcal{X})$ denotes the set of types of n-dimensional sequences and that $\|\cdot\|$ denotes the Euclidean distance). For all sufficiently large γ, $\bar{P}^{(1)}(\gamma)$ and $\bar{P}^{(3)}(\gamma)$ are legitimate probability distributions, and (90) implies that $P^{(2)}(w)$ is a valid probability distribution as well. The probability distribution P_{X^∞} is specified as follows. We let the distribution $P_{X^{\beta_-}}$ of X^{β_-} be uniform over the set of all codewords of type $P^{(1)}(\gamma)$. For $t \in [\beta_- + 1, \beta_+]$, the RVs $\{X_t\}$ are generated independently according to $P^{(2)}(w(t))$ and, for $t \geq \beta_+ + 1$, the RVs $\{X_t\}$ are generated independently according to $P^{(3)}(\gamma)$. For notational

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

convenience, the dependency of $P^{(1)}(\gamma)$ and $P^{(3)}(\gamma)$ on γ is omitted in the remainder of the proof.

We need X^{β_-} to be of constant composition because the capacity-achieving input distributions of the components channels $\{W_k\}$ are generally not given by P^*. The above choice of the distribution of X^{β_-} thereby parallels the achievability proof of the asymptotic expansion of the maximum coding rate for compound DMCs for the fixed blocklength case [3].

We note that $\left\|\bar{P}^{(1)} - P^{(1)}\right\|_1 \leq \mathcal{O}(1/\gamma)$ as $\gamma \to \infty$. Furthermore, differentiability of $I_k(\cdot)$ implies that

$$I_k\Big(P^{(1)}\Big) = I_k\Big(\bar{P}^{(1)}\Big) + \mathcal{O}\left(\frac{1}{\gamma}\right). \tag{362}$$

Additionally, since X^{β_-} are of constant composition, we cannot write $i_{P_{X^{\beta_-}},W_k^{\beta_-}}(X^{\beta_-};Y_k^{\beta_-})$ as a sum of independent RVs since $P_{Y^{\beta_-}}$ is not a product distribution. Hence, to dispose of the dependency among the RVs X^{β_-}, we use the inequality [27, Eq. (4.49)]

$$\begin{aligned} &P_{Y_k^{\beta_-}}(\mathbf{y}) \\ &\quad = P_{X^{\beta_-}} W_k^{\beta_-}(\mathbf{y}) \leq |\mathcal{P}_{\beta_-}(\mathcal{X})|(P^{(1)}W_k)^{\beta_-}(\mathbf{y}) \end{aligned} \tag{363}$$

which holds for all $\mathbf{y} \in \mathcal{Y}_k^{\beta_-}$, and the inequality $|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}$ [17, Th. 11.1.1] to conclude that, for all $t \geq \beta_-$,

$$\begin{aligned} &i_{P_{X^t},W_k^t}(x^t;y_k^t) \\ &\quad = \log \frac{W_k^{\beta_-}(y_k^{\beta_-}|x^{\beta_-})}{P_{Y_k^{\beta_-}(y_k^{\beta_-})}} \\ &\qquad + \sum_{n=\beta_-+1}^{t} i_{P_{X_n},W_{k,n}}(x_n;y_{k,n}) \qquad (364) \\ &\quad \geq \log \frac{\prod_{n=1}^{\beta_-} W_k(y_{k,n}|x_n)}{|\mathcal{P}_{\beta_-}(\mathcal{X})|(P^{(1)}W_k)^{\beta_-}(y_k^{\beta_-})} \\ &\qquad + \sum_{n=\beta_-+1}^{t} i_{P_{X_n},W_{k,n}}(x_n;y_{k,n}) \qquad (365) \\ &\quad \geq \sum_{n=1}^{t} i_{P_{X_n},W_k}(x_n;y_{k,n}) - |\mathcal{X}|\log(\beta_-+1). \qquad (366) \end{aligned}$$

It follows from (357) and (366) that

$$\begin{aligned} \mathbb{P}\Big[\max_k \tau_k \leq t\Big] \geq \mathbb{P}\Bigg[\min_k \Bigg\{ &\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n}) \\ &- |\mathcal{X}|\log(\beta_-+1)\Bigg\} \geq \gamma\Bigg]. \end{aligned} \tag{367}$$

We also note that the marginal probability distribution of X_t, for $t \leq \beta_-$, is given by $P^{(1)}$.

Step 2: We shall next estimate the expected value of $\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n})$, which is given by $\sum_{i=1}^{t} I_k(P_{X_i})$. This is needed to lower-bound $\mathbb{P}\Big[\min_k i_{P_{X^t};W_k^t}(X^t,Y_k^t) \geq \gamma\Big]$ using Hoeffding's inequality for $t \leq \beta_-$ and a multivariate version of the Berry-Esseen central limit theorem for the case $t \geq \beta_- + 1$. We first treat the case $t \in [\beta_- + 1, \beta_+]$. We will return to the case $t \in [\beta_+ + 1, \infty)$ shortly. For $t \in [\beta_- + 1, \beta_+]$, we have that

$$\begin{aligned} &\sum_{i=1}^{t} I_k(P_{X_i}) \\ &\quad = \sum_{i=\beta_-+1}^{t} I_k\Big(P^{(2)}(w(i))\Big) + \beta_- I_k\Big(P^{(1)}\Big) \qquad (368) \\ &\quad = \sum_{i=\beta_-+1}^{t} \Big[C - E_k(w(i)) + C\nabla I_k(\bar{\mathbf{v}}'(w(i)))\Big] \\ &\qquad + \beta_- \left(C + \sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(\beta_-))\right) + \mathcal{O}(1). \qquad (369) \end{aligned}$$

Here, in (368) we used that the marginal distribution of P_{X_t} for $t \in [1, \beta_-]$ is given by $P^{(1)}$. To obtain (369), we used (362) and that $\beta_- = \mathcal{O}(\gamma)$. Furthermore, we performed a Taylor-expansion of $I_k(P^{(1)})$ around P^*, and used that $E_k(s) = C - I_k(P^{(2)}(s)) + C\,\nabla I_k(\bar{\mathbf{v}}'(s))$ (recall the definition of $E_k(s)$ in (91) and of $P^{(2)}$ in (360)). We note that $C\nabla I_k(\bar{\mathbf{v}}'(s)) = \nabla I_k(C\bar{\mathbf{v}}'(s))$ because $\nabla I_k(\cdot)$ is linear. To simplify (369), we shall apply the following two asymptotic expansions, which are proven in Appendix VII-A and Appendix VII-B, respectively:

$$\begin{aligned} &C\nabla I_k(\bar{\mathbf{v}}'(w(t))) \\ &= t\sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(t))) - (t-1)\sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(t-1))) \\ &\quad + \mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right) \end{aligned} \tag{370}$$

and

$$\sum_{i=\beta_-+1}^{t} E_k(w(i)) = \sqrt{\frac{\gamma V}{C^3}} \int_{w(\beta_-)}^{w(t)} E_k(s)\mathrm{d}s + \mathcal{O}(\log\gamma) \tag{371}$$

as $\gamma \to \infty$ for all $t \in [\beta_- + 1, \beta_+]$. By substituting (370) and (371) into (369), we obtain (372)–(373), shown in the top of the next page. Here, (373) follows because $(t - \beta_-) \leq (\beta_+ - \beta_-) = \mathcal{O}(\sqrt{\gamma}\log\gamma)$.

We now move to the case $t \in [\beta_+ + 1, \infty)$ for which, proceeding as in (368)–(373), we obtain that

$$\begin{aligned} \sum_{i=1}^{t} I_k(P_{X_i}) &= \beta_+ \left(C + \sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(\beta_+)))\right) \\ &\quad - \sqrt{\frac{\gamma V}{C^3}} \int_{w(\beta_-)}^{w(\beta_+)} E_k(s)\mathrm{d}s \\ &\quad + \sum_{i=\beta_++1}^{t} I_k\Big(P^{(3)}\Big) + \mathcal{O}(\log^2\gamma) \qquad (374) \\ &= t\left(C + \mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right)\right). \qquad (375) \end{aligned}$$

Here, (374) follows from (373) and (375) follows because $\int_{-\infty}^{\infty} E_k(s)\mathrm{d}s < \infty$ (see (92)) and because $I_k\big(P^{(3)}\big) = C + \mathcal{O}(1/\sqrt{\gamma})$. We have also used that $\nabla I_k(\bar{\mathbf{v}}(w(\beta_+))) = \mathcal{O}(\log\gamma)$, which follows because $\bar{\mathbf{v}}'$ is bounded by (90) and from $w(\beta_+) = \mathcal{O}(\log\gamma)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\sum_{i=1}^{t} I_k(P_{X_i}) = \sum_{i=\beta_-+1}^{t} \left(C + i\sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(i))) - (i-1)\sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(i-1))) \right)$$
$$+ \beta_- \left(C + \sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(\beta_-)) \right) - \sqrt{\frac{\gamma V}{C^3}} \int_{w(\beta_-)}^{w(t)} E_k(s)\mathrm{d}s + \mathcal{O}(\log\gamma) + (t-\beta_-)\mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right) \quad (372)$$
$$= t\left(C + \sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(t))) \right) - \sqrt{\frac{\gamma V}{C^3}} \int_{w(\beta_-)}^{w(t)} E_k(s)\mathrm{d}s + \mathcal{O}(\log^2\gamma). \quad (373)$$

Step 3: We now use (373) and (375) to compute the second and the third term in (356). By applying Hoeffding's inequality and then using (375), we demonstrate in Appendix VII-C that

$$\sum_{t=\beta_++1}^{\infty} \left(1 - \mathbb{P}\Big[\max_k \tau_k \le t \Big] \right) = o(1) \quad (376)$$

as $\gamma \to \infty$. Hence, third term in (356) vanishes as $\gamma \to \infty$. Next, we analyze the second term in (356), which require bounds on $\mathbb{P}\Big[\min_k i_{P_{X^t};W_k^t}(X^t, Y_k^t) \ge \gamma \Big]$ for $t \in [\beta_- + 1, \beta_+]$. Let δ be an arbitrary positive constant. In Appendix VII-D, we show using (373) and a multivariate version of the Berry-Esseen central limit theorem for sums of independent RVs [27, Th. 1.8], [28, Th 1.3] that

$$\mathbb{P}\Big[\min_k i_{P_{X^t},W_k^t}(X^t; Y_k^t) \ge \gamma \Big] \ge \prod_k F_{\bar{H}_{\delta,k}}(w(t)) + \mathcal{O}\left(\frac{\log^2\gamma}{\sqrt{\gamma}}\right) \quad (377)$$

as $\gamma \to \infty$. Here, the $\mathcal{O}(\cdot)$ term is uniform in $t \in [\beta_- + 1, \beta_+]$. The RVs $\{\bar{H}_{\delta,k}\}$ have the cumulative distribution function

$$F_{\bar{H}_{\delta,k}}(w) \triangleq \Phi\left(\frac{1}{\varrho_k} \min_{\nu_k \in \{-1,1\}} \frac{w + \nabla I_k(\bar{\mathbf{v}}(w)) - \frac{1}{C}\int_{-\infty}^{w} E_k(s)\mathrm{d}s}{1 - \delta\nu_k} \right). \quad (378)$$

Thus, we have that

$$\sum_{t=\beta_-+1}^{\beta_+} \mathbb{P}\Big[\max_k \tau_k \le t \Big]$$
$$\ge \sum_{t=\beta_-+1}^{\beta_+} \prod_k F_{\bar{H}_{\delta,k}}\left(-\frac{\gamma/C - t}{\sqrt{\gamma V/C^3}} \right) + \mathcal{O}(\log^3\gamma) \quad (379)$$
$$= \int_{\beta_-}^{\beta_+} \prod_k F_{\bar{H}_{\delta,k}}\left(-\frac{\gamma/C - t}{\sqrt{\gamma V/C^3}} \right) \mathrm{d}t + \mathcal{O}(\log^3\gamma) \quad (380)$$
$$= \int_{\beta_-}^{\beta_+} \prod_k \mathbb{P}\left[\frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\bar{H}_{\delta,k} \le t \right] \mathrm{d}t + \mathcal{O}(\log^3\gamma) \quad (381)$$
$$= \int_{\beta_-}^{\beta_+} \mathbb{P}\left[\max_k \left\{ \frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\bar{H}_{\delta,k} \right\} \le t \right] \mathrm{d}t + \mathcal{O}(\log^3\gamma) \quad (382)$$
$$\ge \mathbb{E}\left[\min\left\{ \beta_+ - \beta_-, \beta_+ - \max_k \left\{ \frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\bar{H}_{\delta,k} \right\} \right\} \right] + \mathcal{O}(\log^3\gamma). \quad (383)$$

Here, (379) follows from (358), from (377), and because $(\beta_+ - \beta_-) = \mathcal{O}(\sqrt{\gamma}\log\gamma)$; (380) follows because $F_{\bar{H}_{\delta,k}}(w)$ is a nondecreasing function in w upper-bounded by one; and (383) follows because, for a continuous RV X with probability density function $p_X(x)$ and $b \ge a$, we have that

$$\int_a^b \mathbb{P}[X \le x]\,\mathrm{d}x$$
$$= \int_a^b \int_{-\infty}^{x} p_X(s)\mathrm{d}s\mathrm{d}x \quad (384)$$
$$= \int_a^b \int_{-\infty}^{\infty} p_X(s)\mathbb{1}\{x \ge s\}\,\mathrm{d}s\mathrm{d}x \quad (385)$$
$$= \int_{-\infty}^{\infty} p_X(x) \int_a^b \mathbb{1}\{x \ge s\}\,\mathrm{d}x\mathrm{d}s \quad (386)$$
$$= \int_{-\infty}^{\infty} p_X(s)\min\{b-a, (b-s)^+\}\,\mathrm{d}s \quad (387)$$
$$\ge \mathbb{E}[\min\{b-a, b-X\}]\,. \quad (388)$$

In (386), we used Tonelli's theorem [29, Th. 15.8]. Now, since δ can be chosen arbitrarily small, we obtain from (383) the asymptotic bound

$$\sum_{t=\beta_-+1}^{\beta_+} \mathbb{P}\Big[\max_k \tau_k \le t \Big]$$
$$\ge \mathbb{E}\left[\min\left\{ \beta_+ - \beta_-, \beta_+ - \max_k \left\{ \frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\bar{H}_k \right\} \right\} \right] + o(\sqrt{\gamma}) \quad (389)$$
$$= \beta_+ - \mathbb{E}\left[\max\left\{ \beta_-, \max_k \left\{ \frac{\gamma}{C} + \sqrt{\frac{\gamma V}{C^3}}\bar{H}_k \right\} \right\} \right] + o(\sqrt{\gamma}) \quad (390)$$
$$= \beta_+ - \frac{\gamma}{C} - \sqrt{\frac{\gamma V}{C^3}}\mathbb{E}\Big[\max\Big\{ -\log\gamma, \max_k \bar{H}_k \Big\} \Big] + o(\sqrt{\gamma}) \quad (391)$$
$$= \beta_+ - \frac{\gamma}{C} - \sqrt{\frac{\gamma V}{C^3}}\mathbb{E}\Big[\max_k \bar{H}_k \Big] + o(\sqrt{\gamma}). \quad (392)$$

Recall that the $\{\bar{H}_k\}$ have cumulative distribution function given in (96).

Finally, substituting (376) and (392) in (356), we obtain the desired result (303).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

A. Proof of (370)

We prove (370), for $t \in [\beta_- + 1, \beta_+]$, through the chain of equalities (393)–(397), shown in the top of the next page. Here, (393) follows because $\gamma/C + \sqrt{\gamma V/C^3} w(t) = t$ (see (358)); moreover, (394) follows from a first-order Taylor expansion of $\bar{\mathbf{v}}(\cdot)$ around $w(t)$, and (397) follows because $\bar{\mathbf{v}}'(\cdot)$ is bounded and because $t \in [\beta_- + 1, \beta_+]$ implies that $w(t) = \mathcal{O}(\log \gamma)$.

B. Proof of (371)

For $t \in [\beta_- + 1, \beta_+]$, we obtain (371) through the following steps:

$$\int_{w(\beta_-)}^{w(t)} E_k(s)\mathrm{d}s$$

$$= \sum_{i=\beta_-+1}^{t} \int_{w(i-1)}^{w(i)} E_k(s)\mathrm{d}s \tag{398}$$

$$= \sum_{i=\beta_-+1}^{t} [w(i) - w(i-1)] E_k(s_i) \tag{399}$$

$$= \sum_{i=\beta_-+1}^{t} [w(i) - w(i-1)] \Big(E_k(w(i)) - [w(i) - s_i] E_k'(s_i') \Big) \tag{400}$$

$$= \sum_{i=\beta_-+1}^{t} [w(i) - w(i-1)] E_k(w(i)) - \sum_{i=\beta_-+1}^{t} (w(i) - w(i-1))(w(i) - s_i) E_k'(s_i') \tag{401}$$

$$= \sqrt{\frac{C^3}{\gamma V}} \sum_{i=\beta_-+1}^{t} E_k(w(i)) + \mathcal{O}\Big(\frac{1}{\gamma}\Big) \sum_{i=\beta_-+1}^{t} E_k'(s_i') \tag{402}$$

$$= \sqrt{\frac{C^3}{\gamma V}} \sum_{i=\beta_-+1}^{t} E_k(w(i)) + \mathcal{O}\Big(\frac{\log \gamma}{\sqrt{\gamma}}\Big) \tag{403}$$

as $\gamma \to \infty$. Here, (399) follows from the mean value theorem for definite integrals [29, Th. 7.30] which implies that there exist constants $s_i \in (w(i-1), w(i))$ satisfying (399); the equality (400) follows from the mean value theorem [29, Th. 5.11], which guarantees the existence of constants $s_i' \in (s_i, w(i))$ such that (400) is satisfied; (402) follows because $w(i) - w(i-1) = \sqrt{C^3/(\gamma V)}$; and (403) holds because $t \in [\beta_- + 1, \beta_+]$, which implies that $(t - \beta_-) = \mathcal{O}(\sqrt{\gamma} \log \gamma)$, and because $\{|E_k'(w)|\}$ are bounded (see (93)).

C. Proof of (376)

We shall first apply Hoeffding's inequality to obtain an upper bound on $1 - \mathbb{P}\Big[\max_k \tau_k \le t\Big]$ that holds for $t \in [\beta_+ + 1, \infty)$. To obtain (376), we then sum this upper bound over all integers larger than β_+.

Observe that (375) implies that there exists a constant $c_1 > 0$ such that, for all sufficiently large γ and for all $k \in \mathcal{K}$, we have

$$\sum_{i=1}^{t} I_k(P_{X_i}) - |\mathcal{X}| \log(\beta_- + 1) \ge t \left(C - \frac{c_1 \log \gamma}{\sqrt{\gamma}} \right). \tag{404}$$

Choose an arbitrary $\tilde{\mathbf{x}} \in \mathcal{X}^{\beta_-}$ of type $P^{(1)}$. Then, we proceed with the steps (405)–(409), shown in the top of the next page. Here, (405) follows from (367); (406) follows from the union bound; (407) follows because the distribution of $\sum_{n=1}^{\beta_-} i_{P_{X_n},W_k}(X_n; Y_{k,n})$ depends only on X^{β_-} through its type, since X^{β_-} is of constant composition; (408) follows from Hoeffding's inequality [24, Th. 2], and because $\{i_{P_{X_n},W_k}(X_n; Y_{k,n})\}_{n=1}^{\beta_-}$ are conditionally independent given X^{β_-}; and (409), for sufficiently large γ, holds because $tC > \gamma$ and because of (404). Consequently, we have

$$\sum_{t=\beta_++1}^{\infty} \Big(1 - \mathbb{P}\Big[\max_k \tau_k \le t\Big]\Big)$$

$$\le K \sum_{t=\beta_+}^{\infty} \exp\left(-\mathbb{c} \left(\frac{t(C - c_1 \log(\gamma)/\sqrt{\gamma}) - \gamma}{\sqrt{\gamma}} \right)^2 \right) \tag{410}$$

$$= K \sum_{i=1}^{\infty} \sum_{t=\lfloor \gamma/C + i\sqrt{\gamma V/C^3} \log \gamma \rfloor}^{\lfloor \gamma/C + (i+1)\sqrt{\gamma V/C^3} \log \gamma \rfloor - 1} \exp\left(-\mathbb{c} \left(\frac{t(C - c_1 \log(\gamma)/\sqrt{\gamma}) - \gamma}{\sqrt{\gamma}} \right)^2 \right) \tag{411}$$

$$\le \mathbb{c}\sqrt{\gamma} \log(\gamma) \sum_{i=1}^{\infty} \exp\Bigg(-\mathbb{c} \left[\left(\frac{\sqrt{\gamma}}{C} + i\mathbb{c} \log \gamma \right) \left(C - \frac{c_1 \log \gamma}{\sqrt{\gamma}} \right) - \sqrt{\gamma} \right]^2 \Bigg) \tag{412}$$

$$\le \mathbb{c}\sqrt{\gamma} \log(\gamma) \sum_{i=1}^{\infty} \exp\Big(-\mathbb{c} \left(i \log \gamma - \mathbb{c} \right)^2 \Big) \tag{413}$$

$$\le \mathbb{c}\sqrt{\gamma} \log(\gamma) \sum_{i=1}^{\infty} \exp\Big(-\mathbb{c} \left(i \log \gamma \right)^2 \Big) \tag{414}$$

$$= \mathbb{c}\sqrt{\gamma} \log(\gamma) \sum_{i=1}^{\infty} \exp\big(-\mathbb{c} \log^2 \gamma \big)^{i^2} \tag{415}$$

$$\le \mathbb{c}\sqrt{\gamma} \log(\gamma) \sum_{i=1}^{\infty} \exp\big(-\mathbb{c} \log^2 \gamma \big)^{i} \tag{416}$$

$$= \mathbb{c}\sqrt{\gamma} \log(\gamma) \frac{\exp\big(-\mathbb{c} \log^2 \gamma\big)}{1 - \exp\big(-\mathbb{c} \log^2 \gamma\big)} = o(1) \tag{417}$$

as $\gamma \to \infty$. Here, (410) follows by (409), (412) follows because $\exp\Big(-\mathbb{c}\left((t(C - c_1 \log(\gamma)/\sqrt{\gamma}) - \gamma)/\sqrt{\gamma}\right)^2\Big)$ decreases in t for sufficiently large γ, and both (413) and (414) hold for sufficiently large γ.

D. Proof of (377)

We shall apply a multivariate version of the Berry-Esseen central limit theorem for sums of independent RVs to $\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n; Y_{k,n})$ in (367). To do so, we need to compute the variance of $\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n; Y_{k,n})$. It turns out convenient to define the unconditional information variance

$$U_k(P) \triangleq \mathrm{Var}_{P \times W_k}[i_{P,W_k}(X; Y_k)] \tag{418}$$

and

$$V_k^t \triangleq \frac{1}{t} \left(\beta_- V_k(P^{(1)}) + \sum_{n=\beta_-+1}^{t} U_k(P_{X_n}) \right). \tag{419}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$
\begin{aligned}
&t\sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(t))) - (t-1)\sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(t-1))) \\
&\quad= \nabla I_k\left(\left(\frac{\gamma}{C}+\sqrt{\frac{\gamma V}{C^3}}w(t)\right)\sqrt{\frac{VC}{\gamma}}\bar{\mathbf{v}}(w(t)) - \left(\frac{\gamma}{C}+\sqrt{\frac{\gamma V}{C^3}}w(t)-1\right)\sqrt{\frac{VC}{\gamma}}\bar{\mathbf{v}}\left(w(t)-\sqrt{\frac{C^3}{\gamma V}}\right)\right) && (393)\\
&\quad= \nabla I_k\left(\left(\frac{\gamma}{C}+\sqrt{\frac{\gamma V}{C^3}}w(t)\right)\sqrt{\frac{VC}{\gamma}}\bar{\mathbf{v}}(w(t)) - \left(\frac{\gamma}{C}+\sqrt{\frac{\gamma V}{C^3}}w(t)-1\right)\sqrt{\frac{VC}{\gamma}}\left(\bar{\mathbf{v}}(w(t)) - \sqrt{\frac{C^3}{\gamma V}}\bar{\mathbf{v}}'(w(t))\right)\right) \\
&\qquad+ \mathcal{O}\left(\frac{1}{\sqrt{\gamma}}\right) && (394)\\
&\quad= \nabla I_k\left(\sqrt{\frac{VC}{\gamma}}\bar{\mathbf{v}}(w(t)) + \left(\frac{\gamma}{C}+\sqrt{\frac{\gamma V}{C^3}}w(t)-1\right)\frac{C^2}{\gamma}\bar{\mathbf{v}}'\left(w(t)\right)\right) + \mathcal{O}\left(\frac{1}{\sqrt{\gamma}}\right) && (395)\\
&\quad= \sqrt{\frac{VC}{\gamma}}\nabla I_k(\bar{\mathbf{v}}(w(t))) + \left(\sqrt{\frac{VC}{\gamma}}w(t) - \frac{C^2}{\gamma}\right)\nabla I_k(\bar{\mathbf{v}}'(w(t))) + C\nabla I_k(\bar{\mathbf{v}}'(w(t))) + \mathcal{O}\left(\frac{1}{\sqrt{\gamma}}\right) && (396)\\
&\quad= C\nabla I_k(\bar{\mathbf{v}}'(w(t))) + \mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right). && (397)
\end{aligned}
$$

$$
\begin{aligned}
1-\mathbb{P}\Big[\max_k \tau_k \le t\Big] &\le 1-\mathbb{P}\left[\min_k\left\{\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n}) - |\mathcal{X}|\log(\beta_-+1)\right\}\ge\gamma\right] && (405)\\
&\le \sum_k \mathbb{P}\left[\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n}) - |\mathcal{X}|\log(\beta_-+1)\le\gamma\right] && (406)\\
&= \sum_k \mathbb{P}\left[\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n}) - |\mathcal{X}|\log(\beta_-+1)\le\gamma\,\middle|\,X^{\beta_-}=\tilde{\mathbf{x}}\right] && (407)\\
&\le \sum_k \exp\left(-\mathbb{c}\left(\frac{\sum_{i=1}^{t} I_k(P_{X_i}) - |\mathcal{X}|\log(\beta_-+1)-\gamma}{\sqrt{\gamma}}\right)^2\right) && (408)\\
&\le K\exp\left(-\mathbb{c}\left(\frac{t(C-c_1\log(\gamma)/\sqrt{\gamma})-\gamma}{\sqrt{\gamma}}\right)^2\right). && (409)
\end{aligned}
$$

Although V_k^t depends on γ, we omit denoting this explicitly to make the notation more convenient. Then, for $t\in[\beta_-+1,\beta_+]$, we have that

$$
\begin{aligned}
&\mathrm{Var}\left[\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n})\right] - \sum_{n=\beta_-+1}^{t} U_k(P_{X_n}) \\
&\quad= \mathbb{E}\left[\mathrm{Var}\left[\sum_{n=1}^{\beta_-} i_{P^{(1)},W_k}(X_n;Y_{k,n})\,\middle|\,X^{\beta_-}\right]\right] \\
&\qquad+ \mathrm{Var}\left[\mathbb{E}\left[\sum_{n=1}^{\beta_-} i_{P^{(1)},W_k}(X_n;Y_{k,n})\,\middle|\,X^{\beta_-}\right]\right] && (420)\\
&\quad= \sum_{n=1}^{\beta_-}\mathbb{E}\big[\mathrm{Var}\big[i_{P^{(1)},W_k}(X_n;Y_{k,n})\big|X_n\big]\big] \\
&\qquad+ \mathrm{Var}\left[\sum_{n=1}^{\beta_-}\mathbb{E}\big[i_{P^{(1)},W_k}(X_n;Y_{k,n})|X_n\big]\,\middle|\,X^{\beta_-}\right] && (421)\\
&\quad= \beta_- V_k(P^{(1)}). && (422)
\end{aligned}
$$

Here, (420) follows from the law of total variance and from (418), (421) follows because $\{i_{P^{(1)},W_k}(X_n;Y_{k,n})\}_{n=1}^{\beta_-}$ are conditionally independent given X^{β_-}, and (422) follows since the marginal distribution of X_n, for $n\in[1,\beta_-]$, is given by $P^{(1)}$ and since X^{β_-} is of constant composition.

Recall that $\tilde{\mathbf{x}}\in\mathcal{X}^{\beta_-}$ has the type $P^{(1)}$. By using (366) and by invoking the multivariate version of the Berry-Esseen central limit theorem for sums of independent RVs reported in [27, Th. 1.8] and [28, Th 1.3], we obtain, for $t\in[\beta_-+1,\beta_+]$, the estimate

$$
\begin{aligned}
&\mathbb{P}\left[\min_k i_{P_{X^t},W_k^t}(X^t;Y_k^t)\ge\gamma\right] \\
&\ge \mathbb{P}\left[\min_k\left\{\sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n})\right\}\right. \\
&\qquad\left.\ge\gamma+|\mathcal{X}|\log(\beta_-+1)\right] && (423)
\end{aligned}
$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$= \mathbb{P}\left[\min_k \left\{ \sum_{n=1}^{t} i_{P_{X_n},W_k}(X_n;Y_{k,n}) \right\} \geq \gamma + |\mathcal{X}| \log(\beta_- + 1) \middle| X^{\beta_-} = \tilde{\mathbf{x}} \right] \quad (424)$$

$$\geq \prod_k Q\left(\frac{\gamma + |\mathcal{X}| \log(\beta_- + 1) - \sum_{n=1}^{t} I_k(P_{X_n})}{\sqrt{V_k^t}} \right) + \frac{\mathbb{C}}{\sqrt{\gamma}} \quad (425)$$

$$= \prod_k Q\left(\frac{\gamma - \sum_{n=1}^{t} I_k(P_{X_n})}{\sqrt{V_k^t}} \right) + \mathcal{O}\left(\frac{\log \gamma}{\sqrt{\gamma}} \right). \quad (426)$$

Here, (424) follows because the distribution of $\sum_{n=1}^{\beta_-} i_{P_{X_n},W_k}(X_n;Y_{k,n})$ depends only on X^{β_-} through its type since X^{β_-} is of constant composition and (425) follows, in addition to the central limit theorem, from (419), from (422), because $\{i_{P_{X_n},W_k}(X_n;Y_{k,n})\}_{n\in\{1,\cdots,\beta_-\},k\in\mathcal{K}}$ are conditional independent given X^{β_-} and because the $\{T_k(\cdot)\}$ are uniformly upper-bounded [4, Lem. 46]. Furthermore, we obtained (426) by performing a first-order Taylor expansion of the Q function around $\left(\gamma - \sum_{n=1}^{t} I_k(P_{X_n})\right)/\sqrt{V_k^t}$.

Next, we approximate V_k^t in (426) by V_k defined in (9) in a sense we shall make precise shortly. Recall that δ is an arbitrarily positive constant. Then, for sufficiently large γ, we have

$$\left| \sqrt{\frac{V_k^t}{(\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}))^3}} - \sqrt{\frac{V_k}{C^3}} \right| \leq \sqrt{\frac{V_k}{C^3}}\delta \quad (427)$$

for every $t \in [\beta_- + 1, \beta_+]$ (recall that P_{X_i} and V_k^t both depends on γ). This follows because $U_k(P)$ and $I_k(P)$ are upper-bounded by

$$U_{\max} \triangleq \max_{\substack{P\in\mathcal{P}(\mathcal{X}) \\ k\in\mathcal{K}}} U_k(P) < \infty \quad (428)$$

and by C_k, respectively, and lower-bounded by 0. Hence, we have that

$$\sqrt{\frac{V_k^t}{(\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}))^3}} \geq \frac{t}{\beta_-}\sqrt{\frac{V_k(P^{(1)})}{\left(I_k(P^{(1)}) + (t/\beta_- - 1)C_k\right)^3}} \quad (429)$$

which converges to $\sqrt{\frac{V_k}{C^3}}$ as $\gamma \to \infty$. This convergence follows from (362), because $t/\beta_- \to 1$, and because $\sqrt{V_k(P^{(1)})/I_k(P^{(1)})^3} \to \sqrt{V_k/C^3}$ as $\gamma \to \infty$. Likewise, we have

$$\sqrt{\frac{V_k^t}{(\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}))^3}} \leq \frac{t}{\beta_-}\sqrt{\frac{V_k(P^{(1)}) + (t/\beta_- - 1)U_{\max}}{I_k(P^{(1)})^3}} \quad (430)$$

which also converges to $\sqrt{V_k/C^3}$ as $\gamma \to \infty$. Consequently, these arguments imply that (427) is satisfied for sufficiently large γ. Similarly to the asymptotic analysis of the converse bound in Appendix IV-C, (427) allows us to approximate $\sqrt{V_k^t/(\frac{1}{t}\sum_{i=1}^{t} I_k(P_{X_i}))^3}$ by $\sqrt{V_k/C^3}$ and, hence, eliminate the dependency on γ.

In order to further bound (426) for $t \in [\beta_- + 1, \beta_+]$, we shall now use (427) together with the inequality (proved in Appendix VIII)

$$\frac{\gamma - \xi a}{\sqrt{\xi b}} \leq \frac{\gamma - \xi a}{\sqrt{\gamma b/a}} + \sqrt{\frac{b}{a\gamma}}\left(\frac{\gamma - \xi a}{\sqrt{\xi b}}\right)^2 \quad (431)$$

which holds for all positive a, b, ξ, and γ. This implies the steps (432)–(435), shown in the top of this page. Here, (433) holds for sufficiently large γ, (434) follows from (427), and (435) follows because the derivative of the Q function is bounded. Finally, we substitute (373) into (435) and get the chain of inequalities (436)–(439), shown in the top of the next page. Here, (436) follows because $\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}) = C + \mathcal{O}(1/\sqrt{\gamma})$. Furthermore, we applied a Taylor expansion of γ/x around $x = C$, which implies that there exists a positive constant c_2 such that the following inequality holds for all sufficiently large γ:

$$\frac{\gamma}{\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n})} \leq \frac{\gamma}{C} - \frac{\gamma}{C^2}\left(\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}) - C\right) + c_2\gamma\left(\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}) - C\right)^2 \quad (440)$$

$$\leq \frac{\gamma}{C} - \frac{\gamma}{C^2}\left(\frac{1}{t}\sum_{n=1}^{t} I_k(P_{X_n}) - C\right) + \mathbb{c}. \quad (441)$$

Moreover, (437) follows from (373), and (438) follows because the derivative of the Q-function is bounded and because

$$\sqrt{\gamma}\left|\frac{1}{C} - \frac{\gamma}{tC^2}\right| = \mathcal{O}(\log\gamma). \quad (442)$$

To prove (442), we used that $t \in [\beta_- + 1, \beta_+]$ and we applied a first-order Taylor expansion of the Q function, and absorbed the remainder term in the $\mathcal{O}\big(\log^2(\gamma)/\sqrt{\gamma}\big)$ term. Finally, (439) follows from (371) and by the definition of the cumulative distribution functions of the RVs $\{\bar{H}_{\delta,k}\}$:

$$F_{\bar{H}_{\delta,k}}(w) \triangleq \Phi\left(\frac{1}{\varrho_k}\min_{\nu_k\in\{-1,1\}} \frac{w + \nabla I_k(\bar{\mathbf{v}}(w)) - \frac{1}{C}\int_{-\infty}^{w} E_k(s)\mathrm{d}s}{1 - \delta_2\nu_k}\right). \quad (443)$$

Appendix VIII
Basic Lemmas

Lemma 13: Fix arbitrary $x \in \mathbb{R}$, $a > 0, b > 0$, and $\lambda > 0$. Suppose that ξ is the unique solution to the equation

$$\frac{\lambda - \xi a}{\sqrt{b\xi}} = x. \quad (444)$$

Then we have:

$$0 \leq \xi - \left(\frac{\lambda}{a} - x\sqrt{\frac{\lambda b}{a^3}}\right) \leq \frac{b}{a^2}x^2. \quad (445)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

$$\mathbb{P}\left[\min_k i_{P_{X^t},W_k^t}(X^t;Y_k^t)\geq\gamma\right]$$
$$\geq\prod_k Q\left(\frac{\gamma-\sum_{n=1}^t I_k(P_{X_n})}{\sqrt{\gamma V_k^t/\left(\frac{1}{t}\sum_{n=1}^t I_k(P_{X_n})\right)}}+\sqrt{\frac{V_k^t}{\gamma\sum_{n=1}^t I_k(P_{X_n})}}\left(\frac{\gamma-\sum_{n=1}^t I_k(P_{X_n})}{\sqrt{V_k^t}}\right)^2\right)+\mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right) \tag{432}$$
$$\geq\prod_k Q\left(\frac{\gamma/\left(\frac{1}{t}\sum_{n=1}^t I_k(P_{X_n})\right)-t}{\sqrt{\gamma V_k^t/\left(\frac{1}{t}\sum_{n=1}^t I_k(P_{X_n})\right)^3}}+\mathbb{c}\frac{\log^2\gamma}{\sqrt{\gamma}}\right)+\mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right) \tag{433}$$
$$\geq\prod_k Q\left(\max_{\nu_k\in\{-1,1\}}\frac{\gamma/\left(\frac{1}{t}\sum_{n=1}^t I_k(P_{X_n})\right)-t}{\sqrt{\gamma V_k/C^3}(1-\delta_2\nu_k)}+\mathbb{c}\frac{\log^2\gamma}{\sqrt{\gamma}}\right)+\mathcal{O}\left(\frac{\log\gamma}{\sqrt{\gamma}}\right) \tag{434}$$
$$\geq\prod_k Q\left(\max_{\nu_k\in\{-1,1\}}\frac{\gamma/\left(\frac{1}{t}\sum_{n=1}^t I_k(P_{X_n})\right)-t}{\sqrt{\gamma V_k/C^3}(1-\delta_2\nu_k)}\right)+\mathcal{O}\left(\frac{\log^2\gamma}{\sqrt{\gamma}}\right). \tag{435}$$

$$\mathbb{P}\left[\min_k i_{P_{X^t},W_k}(X^t;Y_k^t)\geq\gamma\right]$$
$$\geq\prod_k Q\left(\max_{\nu_k\in\{-1,1\}}\frac{\frac{\gamma}{C}-\frac{\gamma}{C^2}\left(\frac{1}{t}\sum_{n=1}^t I_k(P_{X_n})-C\right)+\mathbb{c}-t}{\sqrt{\gamma V_k/C^3}(1-\delta_2\nu_k)}\right)+\mathcal{O}\left(\frac{\log^2\gamma}{\sqrt{\gamma}}\right) \tag{436}$$
$$\geq\prod_k Q\left(\max_{\nu_k\in\{-1,1\}}\frac{\frac{\gamma}{C}-\sqrt{\frac{\gamma V}{C^3}}\nabla I_k(\bar{\mathbf{v}}(w(t)))+\frac{\gamma}{tC^2}\sqrt{\frac{\gamma V}{C^3}}\int_{w(\beta_-)}^{w(t)}E_k(s)\mathrm{d}s+\mathbb{c}\log^2\gamma-t}{\sqrt{\gamma V_k/C^3}(1-\delta_2\nu_k)}\right)+\mathcal{O}\left(\frac{\log^2\gamma}{\sqrt{\gamma}}\right) \tag{437}$$
$$\geq\prod_k Q\left(\frac{1}{\varrho_k}\max_{\nu_k\in\{-1,1\}}\frac{-w(t)-\nabla I_k(\bar{\mathbf{v}}(w(t)))+\frac{1}{C}\int_{-\infty}^{w(t)}E_k(s)\mathrm{d}s}{1-\delta_2\nu_k}\right)+\mathcal{O}\left(\frac{\log^2\gamma}{\sqrt{\gamma}}\right) \tag{438}$$
$$\geq\prod_k F_{\bar{H}_{\delta,k}}(w(t))+\mathcal{O}\left(\frac{\log^2\gamma}{\sqrt{\gamma}}\right). \tag{439}$$

The inequalities in (445) are equivalent to

$$\frac{\lambda-\xi a}{\sqrt{\lambda b/a}}\leq x\leq\frac{\lambda-\xi a}{\sqrt{\lambda b/a}}+\sqrt{\frac{b}{a\lambda}}x^2. \tag{446}$$

Proof: For all $x\in\mathbb{R}$, we have that

$$\xi=\frac{\lambda}{a}+\frac{b}{2a^2}x^2-x\sqrt{\frac{b^2}{4a^4}x^2+\frac{b\lambda}{a^3}}. \tag{447}$$

When $x\geq 0$,

$$\frac{\lambda}{a}-x\sqrt{\frac{b\lambda}{a^3}}\leq\frac{\lambda}{a}+\frac{b}{2a^2}x^2-x\sqrt{\frac{b^2}{4a^4}x^2+\frac{b\lambda}{a^3}}$$
$$=\xi \tag{448}$$
$$\leq\frac{\lambda}{a}+\frac{b}{2a^2}x^2-x\sqrt{\frac{b\lambda}{a^3}}. \tag{449}$$

Furthermore, when $x\leq 0$,

$$\frac{\lambda}{a}+\frac{b}{2a^2}x^2-x\sqrt{\frac{b\lambda}{a^3}}\leq\frac{\lambda}{a}+\frac{b}{2a^2}x^2-x\sqrt{\frac{b^2}{4a^4}x^2+\frac{b\lambda}{a^3}} \tag{450}$$
$$=\xi \tag{451}$$
$$\leq\frac{\lambda}{a}+\frac{b}{a^2}x^2-x\sqrt{\frac{b\lambda}{a^3}}. \tag{452}$$

This establishes (446). ∎

Lemma 14: Fix an integer $K\geq 2$. Let $\{x_j\}_{j=1}^{K-1}$ be constants such that

$$\sum_{j=1}^i x_j>0\qquad\text{for } i\in\{1,\cdots,K-1\}. \tag{453}$$

Then, there exist positive constants $\{\zeta_i\}_{i=1}^{K-2}$ such that

$$x_i+\zeta_{i-1}-\zeta_i>0\qquad\text{for } i\in\{1,\cdots,K-1\}. \tag{454}$$

In (454), we set $\zeta_0\triangleq\zeta_{K-1}\triangleq 0$.

Proof: The lemma is obviously satisfied when $K=2$. Next, we consider the case $K\geq 3$. Define

$$\delta\triangleq\min_{i\in\{1,\cdots,K-1\}}\frac{x_1+\cdots+x_i}{K-1} \tag{455}$$

and let $\zeta_i\triangleq x_1+\cdots+x_i-i\delta$ for $i\in\{1,\cdots,K-2\}$. Note that ζ_i is positive for $i\in\{1,\cdots,K-2\}$. Then, we establish (454) for $i\in\{1,\cdots,K-2\}$ as follows

$$x_i+\zeta_{i-1}-\zeta_i$$
$$=x_i+(x_1+\cdots+x_i-(i-1)\delta)$$
$$\quad-(x_1+\cdots+x_i-i\delta) \tag{456}$$
$$=\delta \tag{457}$$
$$>0. \tag{458}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2868953, IEEE Transactions on Information Theory

Here, (458) follows from (453) and (455). To prove (454) for $i = K-1$, we proceed as follows

$$\begin{aligned} & x_{K-1} + \zeta_{K-2} - \zeta_{K-1} \\ & = x_1 + \cdots + x_{K-1} - (K-2)\delta && (459) \\ & \geq x_1 + \cdots + x_{K-1} - \frac{K-2}{K-1}(x_1 + \cdots + x_{K-1}) && (460) \\ & > 0. && (461) \end{aligned}$$

■

Lemma 15: Define $\psi(x) \triangleq \phi(x)/\Phi(x)$. Then the following holds:

a) $\psi'(x) \in (-1, 0)$ for all $x \in \mathbb{R}$,
b) $\psi''(x)$ is positive for all $x \in \mathbb{R}$,
c) $\beta\psi'(\psi^{-1}(x)) < \psi'(\psi^{-1}(\beta x))$ for all $x > 0$ and $\beta > 1$.

Proof: Define $\nu(x) \triangleq \psi(-x)$. Then,

$$\frac{1}{\nu(x)} = e^{\frac{1}{2}x^2} \int_x^\infty e^{-\frac{1}{2}u^2} \mathrm{d}u. \tag{462}$$

This quantity is known as Mill's ratio [30]. It follows from [30, Eq. (3)] that $\nu'(x) \in (0, 1)$, which implies that $\psi'(x) \in (-1, 0)$. Similarly, (b) follows from [30, Eq. (4)], which states that $\nu''(x) > 0$, thereby implying that $\psi''(x) > 0$.

To establish (c), we use the identity

$$\begin{aligned} \psi'(\psi^{-1}(x)) &= -\psi(\psi^{-1}(x))(\psi(\psi^{-1}(x)) + \psi^{-1}(x)) && (463) \\ &= -x(x + \psi^{-1}(x)). && (464) \end{aligned}$$

This implies that

$$\begin{aligned} & \psi'(\psi^{-1}(\beta x)) - \beta\psi'(\psi^{-1}(x)) \\ & = -\beta x(\beta x + \psi^{-1}(\beta x)) + \beta x(x + \psi^{-1}(x))) && (465) \\ & = \beta x \left[x + \psi^{-1}(x) - \beta x - \psi^{-1}(\beta x)\right] && (466) \\ & > 0. && (467) \end{aligned}$$

Here, (465) follows from (464) and (467) follows because $x + \psi^{-1}(x)$ is a decreasing function. ■

REFERENCES

[1] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Variable-length coding with stop-feedback for the common-message broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016.

[2] A. El Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge Univ. Press, 2011.

[3] Y. Polyanskiy, "On dispersion of compound DMCs," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, USA, 2013, pp. 26–32.

[4] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[5] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 10–30, Oct-Dec 1976.

[6] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 25, no. 6, pp. 729–733, Nov. 1979.

[7] P. Berlin, B. Nakiboglu, B. Rimoldi, and E. Teletar, "A simple converse of Burnashev's reliability function," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3074–3080, Jul. 2009.

[8] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[9] Y. Altug and A. B. Wagner, "Feedback can improve the second-order coding performance in discrete memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, Jul. 2014.

[10] A. Tchamkerten and E. Teletar, "A feedback strategy for binary symmetric channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Lausanne, Switzerland, Jul. 2002, p. 362.

[11] ——, "Variable length coding over an unknown channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.

[12] R. Devassy, G. Durisi, B. Lindqvist, W. Yang, and M. Dalai, "Nonasymptotic coding-rate bounds for binary erasure channels with feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, United Kingdom, sep 2016.

[13] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Broadcasting a common message with variable-length stop-feedback codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 2015, pp. 2505–2509.

[14] P. Billingsley, *Probability and Measure, Anniversary Ed.* Hoboken, NJ, USA: Wiley, 2012.

[15] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[16] H. G. Eggleston, *Convexity*. New York, NY, USA: Cambridge Univ. Press, 2009.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory, 2nd ed.* Hoboken, NJ, USA: Wiley Interescience, 2012.

[18] H. R. Lerche, *Boundary Crossing of Brownian Motion - Its Relation to the Law of the Iterated Logarithm and to Sequential Analysis*. Berlin, Germany: Springer-Verlag, 1986.

[19] G. Deelstra, "Remarks on 'Boundary crossing results for Brownian motion'," *Blätter der DGVFM*, pp. 449–456, Oct. 1994.

[20] R. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ, USA: Wiley, 1968.

[21] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Common-message broadcast channels with feedback in the nonasymptotic regime: Full feedback," *IEEE Trans. Inf. Theory*, Jul. 2018, to appear.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[23] I. Csiszár and J. Körner, *Information Theory: Coding Theorem for Discrete Memoryless Systems, 2nd ed.* New York, NY, USA: Cambridge Univ. Press, 2012.

[24] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

[25] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7041–7051, Nov. 2013.

[26] V. V. Petrov, *Sums of Independent Random Variables*. Berlin, Germany: Springer, 1975, translated from the Russian by A. A. Brown.

[27] V. Y. F. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Now Publisher*, vol. 10, no. 4, pp. 1–184, 2014.

[28] F. Götze, "On the rate of convergence in the multivariate CLT," *Ann. Prob.*, vol. 19, no. 2, pp. 724–739, 1991.

[29] T. M. Apostel, *Mathematical Analysis, 2nd ed.* Reading, MA, USA: Addison-Wesley Publishing Company, 1974.

[30] M. R. Sampford, "Some inequalities on Mills' ratio and related functions," *Ann. Math. Stat.*, vol. 24, no. 1, pp. 132–134, 1953.

Kasper Fløe Trillingsgaard (S'12) received his B.Sc. degree in electrical engineering, his M.Sc. degree in wireless communications, and his Ph.D. degree in electrical engineering from Aalborg University, Denmark, in 2011, 2013, and 2017, respectively. He is currently a postdoctoral researcher at the same institution. He was a visiting student at New Jersey Institute of Technology, NJ, USA, in 2012 and at Chalmers University of Technology, Sweden, in 2014. His research interests are in the areas of information and communication theory.

Wei Yang (S'09–M'15) received the B.E. degree in communication engineering and M.E. degree in communication and information systems from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008 and 2011, and the Ph.D. degree in Electrical Engineering from Chalmers University of Technology, Gothenburg, Sweden, in 2015. In the summers of 2012 and 2014, he was a visiting student at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA. From 2015 to 2017, he was a postdoctoral research associate at Princeton University, Princeton, NJ. In Sep. 2017, he joined Qualcomm Research, San Diego, CA, where he is now a senior engineer.

Giuseppe Durisi (S'02–M'06–SM'12) received the Laurea degree summa cum laude and the Doctor degree both from Politecnico di Torino, Italy, in 2001 and 2006, respectively. From 2006 to 2010 he was a postdoctoral researcher at ETH Zurich, Zurich, Switzerland. In 2010, he joined Chalmers University of Technology, Gothenburg, Sweden, where he is now professor and co-director of Chalmers information and communication technology Area of Advance.

Dr. Durisi is a senior member of the IEEE. He is the recipient of the 2013 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East, and Africa Region, and is co-author of a paper that won a "student paper award" at the 2012 International Symposium on Information Theory, and of a paper that won the 2013 IEEE Sweden VT-COM-IT joint chapter best student conference paper award. In 2015, he joined the editorial board of the IEEE TRANSACTIONS ON COMMUNICATIONS as associate editor. From 2011 to 2014, he served as publications editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. His research interests are in the areas of communication theory, information theory, and machine learning.

Petar Popovski (S'97–A'98–M'04–SM'10–F'16) is a Professor of Wireless Communications with Aalborg University. He received the Dipl. Ing. degree in electrical engineering and the Magister Ing. degree in communication engineering from the "Sts. Cyril and Methodius" University, Skopje, Republic of Macedonia, in 1997 and 2000, respectively, and the Ph.D. degree from Aalborg University, Denmark, in 2004. He has over 300 publications in journals, conference proceedings, and edited books. He holds over 30 patents and patent applications. He received an ERC Consolidator Grant (2015), the Danish Elite Researcher award (2016), the IEEE Fred W. Ellersick prize (2016), and the IEEE Stephen O. Rice prize (2018). He is currently a Steering Committee Member of IEEE SmartGridComm and previously served as a Steering Committee Member of the IEEE INTERNET OF THINGS JOURNAL. He is also an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. His research interests are in the area of wireless communication and networking, and communication/information theory.